9-PRISSMA Project
Plateforme de Recherche et d'Investissementpour la Sûreté et la Sécurité de la Mobilité Autonome 04/2021 - 04/2024

[L6.3] WP6 STATE OF THE ART RISK ASSESSMENT AND CERTIFICATION FOR AI: FINAL REPORT

RAPPORT FINAL : ÉTAT DE L'ART EVALUATION DES RISQUES ET CERTIFICATION DE L'IA

Main authors: H-M. HUSSAIN and D. CARITEY (Airbus Protect) and V. BARBOSA (LNE)

Keywords: Risk Assessment, Safety, Certification, AI

Abstract. This document aims to establish a state of the art of risk assessments in different sectors and initiates mapping certification efforts to evaluate artificial intelligence. The first part of the document focuses on the state of the art of risk assessment in transportation means such as aeronautics and Rail, in critical infrastructures such as off-shore and nuclear plants and finally in medical robots as they perform automated tasks in close interface with human beings. The second part presents a mapping of certification, labeling standards for AI evaluation as well as an example of certification of processes for AI.

Résumé. Ce document a pour but d'établir un état de l'art de différentes méthodes d'analyse de risque et initie une cartographie des efforts de certification pour évaluer l'intelligence artificielle. La première partie établit un état d'art dans les domaines du transport tels que l'aéronautique et le ferroviaire, les infrastructures critiques dans l'industrie nucléaire et pétrolière et enfin la robotique médicale qui réalise des actions automatisées en proximité étroite ou en interface des êtres humains. La deuxième partie présente une cartographie des standards de certification et labellisation pour l'évaluation de l'IA ainsi qu'un exemple de processus de certification pour l'IA.

Table of content

1. Introduction

The aim of the deliverable is to identify the references in operation concerning embedded systems and mobility of which automobile is a major actor.

After the first year of state-of-the-art analysis, an incremental surveillance of state of-practices and R&D concerning the use of Artificial Intelligence (AI) for autonomous vehicles, as well as an evaluation of the safety of systems containing AI, has been carried out until the end of the project.

A part of this work is jointly done with the deliverable 1.3 of the PRISSMA project, but the identification of exiting references in this deliverable could concern other sectors than automobile such as aeronautics, robotics, medical, rail and any other fields for which are already integrating AI in their process of system engineering or in the process to integrate it.

This document is an update of the deliverable 6.2

2. State of the art of safety evaluation references

This chapter details all the safety evaluation systems that different sectors have put in place in order to guarantee the safe operation of a system (or plant) in his context.

2.1. Rail

Since the beginning, Railway has a culture of safety and processes to perform follow up and controls that are particularly strict. Initially based on equipment reliability and respect of regulations, these processes now encompass all the principles of Management of safety.

Basic rail regulations widely follow NLF. A series of directives provides the essential guildelines, completed by individual standards. Member states ensure the compliance. Definition of safety requirements depends on the concerned subsystem (mobile, trackside). Mobile subsystem include: rolling stock, on-board control command and signalling.

2.1.1. Rules and Regulations

The modalities regarding the approval for trackside or on-board railway products derive mainly from the Implementing Regulation (EU) 2018/545 and the Directive (EU) 2016/797 on the interoperability of railway systems in the EU (TSI). This latter prescribes requirements for all parts of the railway system as well as the form of proof of conformity with them. The Directive (EU) 2016/797 also contains essential requirements that subsystems must fulfil (Annex III to Directive (EU) 2016/797). Before a mobile subsystem - ultimately a train - can be put into operation, an authorization for placing in service is required. This must be applied for at the European Railway Agency (ERA).

In order to obtain authorization for placing in service, compliance with the essential requirements must be demonstrated in accordance with Art. 15 of Directive (EU) 2016/797. In addition, according to Art. 13 (3) Implementing Regulation (EU) 2018/545, proof of safety and proof of implementation of a risk management process according to Regulation (EU) No. 402/2013 (for aspects not covered by TSIs or national regulations) is required; it must be demonstrated that the various components have been safely integrated.

The basis for verifying this safety is formed by the railway-specific standards DIN EN 50126, 50128, 50129 and the more general standards, e.g. IEC 61508, DIN EN 62061, ISO 26262 and, where applicable, DIN EN 62443, provided they are relevant to the subsystem in question.

There specific functions are explained below:

- **EN 50126** related to the specification and demonstration of reliability, availability, maintainability and Safety (RAMS);
- **EN 50128** related to communication, signaling and processing systems-software for railway control and protection systems;
- **EN 50129** also related to communication, signaling and processing systems but to safety related electronic systems for signaling; The structure of the safety proof is defined in section 7 of DIN EN 50129.

2.1.2. Safety analysis process

All critical systems have to verify the requirements specified in the associated norms.

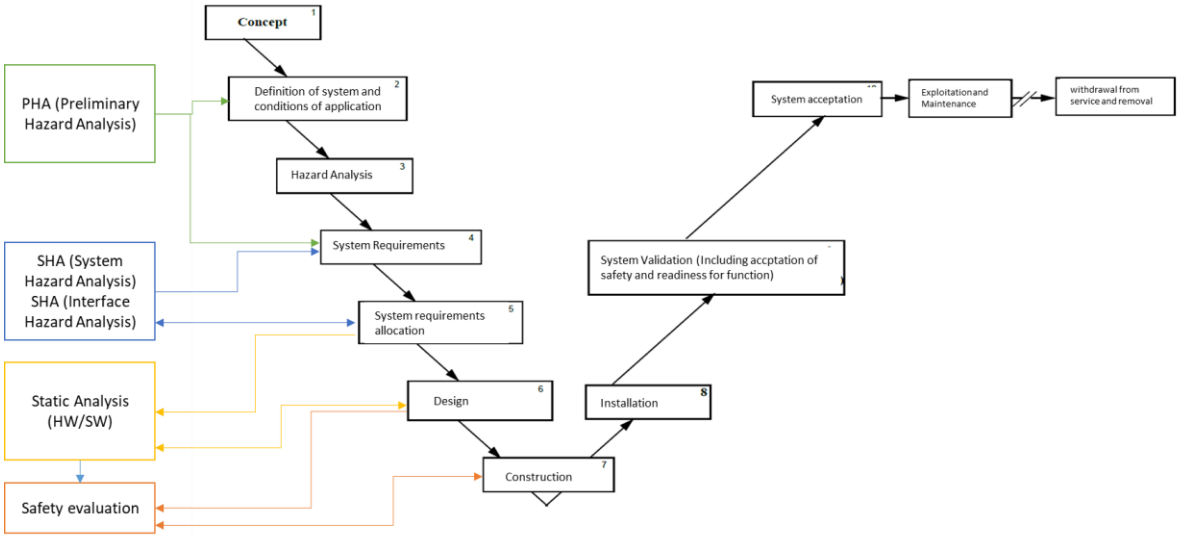

Figure 1: Safety management process integrated to Rail V cycle

Safety analysis are carried out by specialist teams which are independent from design teams.
The global safety process is unfolded through 5 principle phases:

1. A definition of system that identifies whether the change is impacting safety;
2. A Preliminary Risk or Hazard Analysis;
3. Acceptance of risk following a non-regressive approach;
4. Definition of safety requirements;
5. Demonstration of compliance to these requirements by the follow ups and tests.

2.1.2.1. **A definition of system that identifies whether the change is impacting safety**

When a proposed change has an impact on safety, the Common Safety Method for Risk Evaluation and Assessment (CSM RA) requires the proposer to decide, by expert judgement, the significance of the change based on stated criteria (Article 4 [1]).
These criteria are:

- failure consequence: credible worst-case scenario in the event of failure of the system under assessment, taking into account the existence of safety barriers outside the system;
- novelty used in implementing the change: this concerns both what is innovative in the railway sector, and what is new just for the organization implementing the change;
- complexity of the change;
- monitoring: the inability to monitor the implemented change throughout the system life-cycle and take appropriate interventions;
- reversibility: the inability to revert to the system before the change;
- additionality: assessment of the significance of the change taking into account all recent safety-related modifications to the system under assessment and which were not judged as significant.

The CSM RA does not prescribe how to use the criteria, or the priority or weighting given to any of them.

2.1.2.2. **A preliminary risk or hazard analysis**

Preliminary Risk Analysis (PRA)/ Preliminary Hazard Analysis (PHA) can be either performed according to deductive approach going from failures events to digging in the causes or inductive with identifying causes and their consequences.

The deductive approach allows:

- Identification of the risks;
- Allocation of safety requirements (occurrence rate linked to criticality);
- Mitigation of risks;
- Building of the Register of Dangerous Situations (RSD);
- Confrontation of analysis results with safety objectives.

The inductive approach allows, stemming from the general serious dangers in rail classified by time (phase) and space (station, in transit, within the train our out of train, etc.) to reach to reasons that could cause them and define coverage characteristics by sub-systems.

2.1.2.3. **Acceptance of risk following a non-regressive approach**

For a non-innovative system identifying deviations regarding to a referential is sufficient. Indeed, following the GAME [1] approach the safety analysis can focus on deviation and associated analysis.

For innovative systems, the safety study must be standalone and must contain all the justification from PRA to RSD.

2.1.2.4. **Definition of safety requirements**

Rail standard EN 50126 introduces Safety Integrity Levels (SIL). The levels correspond to sets of safety and integrity requirements according to risks.

- SIL 1 – 2: risk of casualty
- SIL 3 – 4: risk of at least on fatality.

According to standard EN 50129 related to rail electronic systems the failures rates per hour must be comprised in a Tolerable Hazard Rate (THR) range for each SIL.

SIL	THR
1	$10^{-6} \leq THR < 10^{-5}$
2	$10^{-7} \leq THR < 10^{-6}$
3	$10^{-8} \leq THR < 10^{-7}$
4	$10^{-9} \leq THR < 10^{-8}$

Table 1: THR associated to SIL

According to EN50128, the allocation of safety functions to the software and its interfaces shall be identified in the documentation of the system. The host system in which the software will be integrated shall be completely defined in terms of:

- Functions and interfaces;
- Operational conditions;

[1] Globalement Au Moins Equivalent (GAME) : At least équivalent

- Configuration or system architecture;
- Hazardous situation to comply;
- Integrity and safety requirements;
- Allocation of SIL level requirements to the hardware and software;
- Time constraints.

For software, SIL can vary from 0 (the lowest) to SIL 4 (the highest). The Safety and Integrity Level of the software shall be decided and allocated with regards to the SIL of the system and the associated risk with the use of the software.

The table below is an extract from Annex A of EN50128. It shows according to SIL what analysis activity should be conducted. R is for Recommended activity. HR is for Highly Recommended.

Method/measure	SIL 0	SIL 1	SIL 2	SIL 3	SIL 4
Static software analysis	R	HR	HR	HR	HR
Dynamic software analysis	-	R	R	HR	HR
Cause/consequence diagrams	R	R	R	R	R
Event Tree analysis	-	R	R	R	R
Analysis of consequences of software errors	-	R	R	HR	HR

Table 2: Software analysis method

2.1.2.5. **Demonstration of compliance to these requirements by the follow ups and tests**

There are two major reasons of verification:

- To ensure that rail system is adapted for operating conditions such as weather or electrical conditions by performing tests of identification and qualification;
- To comply with system requirements by building design justification reports, architecture and interfaces documentation.

Compliance report of meeting the safety requirements is reported in three stages:

- Performing a FMECA on elements of rail systems demonstrating the ability to meet the safety requirements;
- Deducing specific safety tests to be performed;
- Verifying the behavior vs predicted behavior during tests.

In order to guarantee the same safety level across the European rail network, the European directive 2004/49/CE defined a Common Safety Method (CSM). The Safety Management System (SMS) clarifies the accountability of each actor taking interfaces into account. It allows to harmonize the same tools like CSM, Common Safety Indicators (CSI) or Common Safety Objectives (CSO) to assess safety and performances of each actors.

The figure below describes the Common Safety Method.

1. The two first steps “Significant change” and “System definition” correspond to system definition;
2. The following steps in the green bloc correspond to identification of dangers, classification of dangers and assessing either it is acceptable or not;
3. The next steps (in brown) describe three streams of principles in order to accept the dangers;
4. The steps colored in blue correspond to risk evaluation;
5. The second last step corresponds to definition of safety requirements;
6. The final step aims to demonstrate the compliance to the requirements.

The risk assessment can affect the definition of the system since a noncompliance to safety requirements can drive a review of definition of system.

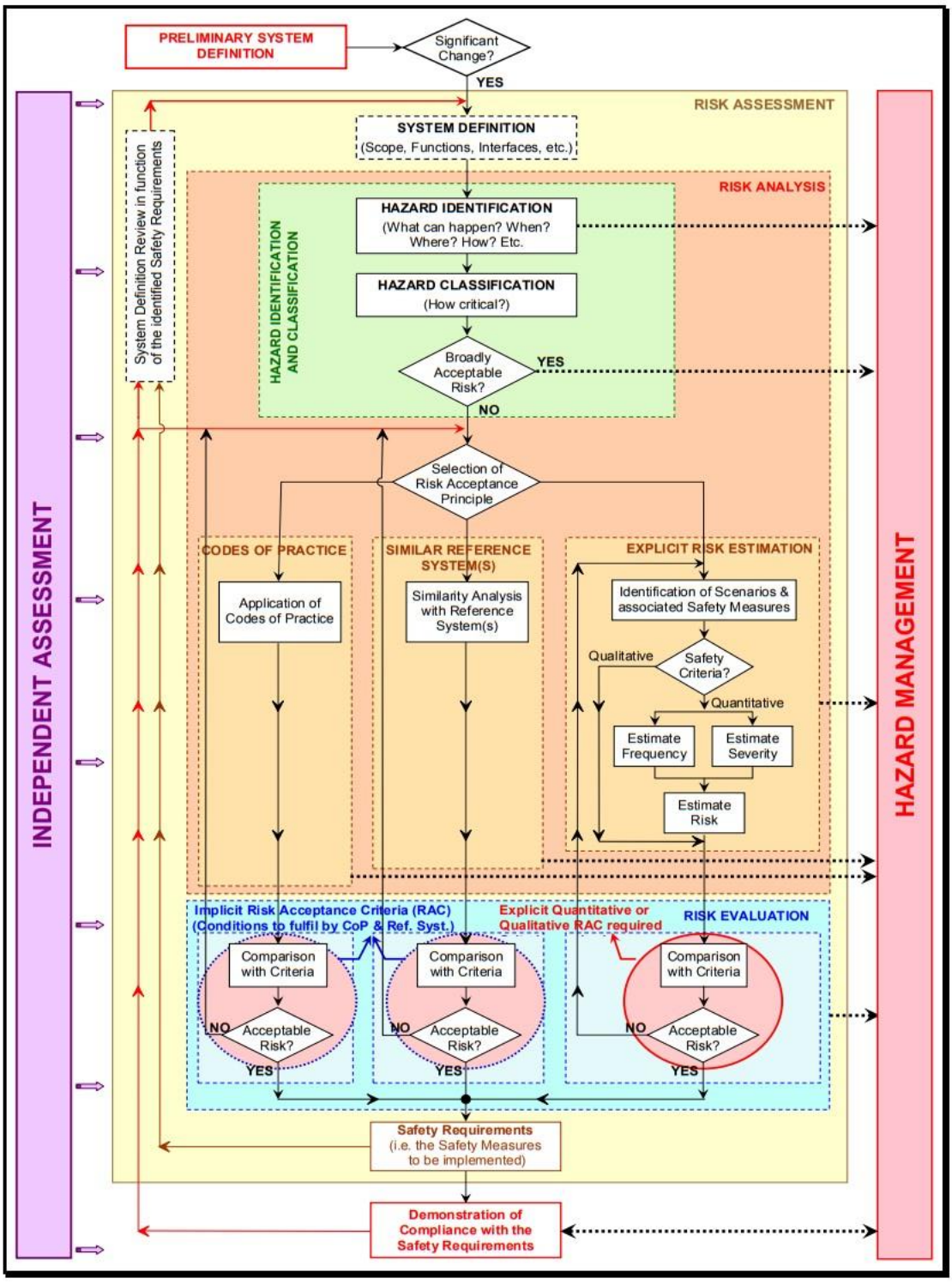

Figure 2: Risk management framework in the CSM Regulation (from Guide for the application of the Common Safety Methods on risk assessment)

2.1.3. Methods and tools used for safety analysis

The first method is to clearly define the requirement necessary for a system beforehand its design and manage them all along its lifecycle.

To express clearly the requirements and the primary design of the systems SysML a UML graphic chart could be used. To quantify the reliability of a system during its life cycle, fault tree analysis editors such as Fault Tree, Risk Spectrum, GRIF or others should be used.

Other techniques, recommended by standards, use formal methods like modelzation by Place/transition net or method B.That is where Model Base Safety Analysis intervene. Tools like SIMFIANEO or MEFISO could be useful.

2.2. Aeronautics

Travel by airplane is the safest mean of transport. As the hazards can have very serious consequences, the aeronautics industry follows stringent safety rules.

2.2.1. Rules and Regulations

This part is covered by deliverable 1.3 wich is a state of the art on AI evaluation, certification and homologation.

The two standards which form the basis to ensure safety are reminded in this paragraph. They are issued by the Society of Automotive Engineers (SAE) and the International Civil Aviation Organization (ICAO).

- **ARP4761/ED135**: Aerospace Recommended Practice Guidelines and Methods for Conducting the Safety Assessment Process on Civil Airborne Systems and Equipment
- **ARP4754/ED79**: Aerospace Recommended Practice Guidelines for Development of Civil Aircraft and Systems

2.2.2. Safety analysis process

The ARP4761 [2] and the ARP4754 [3] describe how the safety analysis is unfolded. It consists of three categories of analysis:

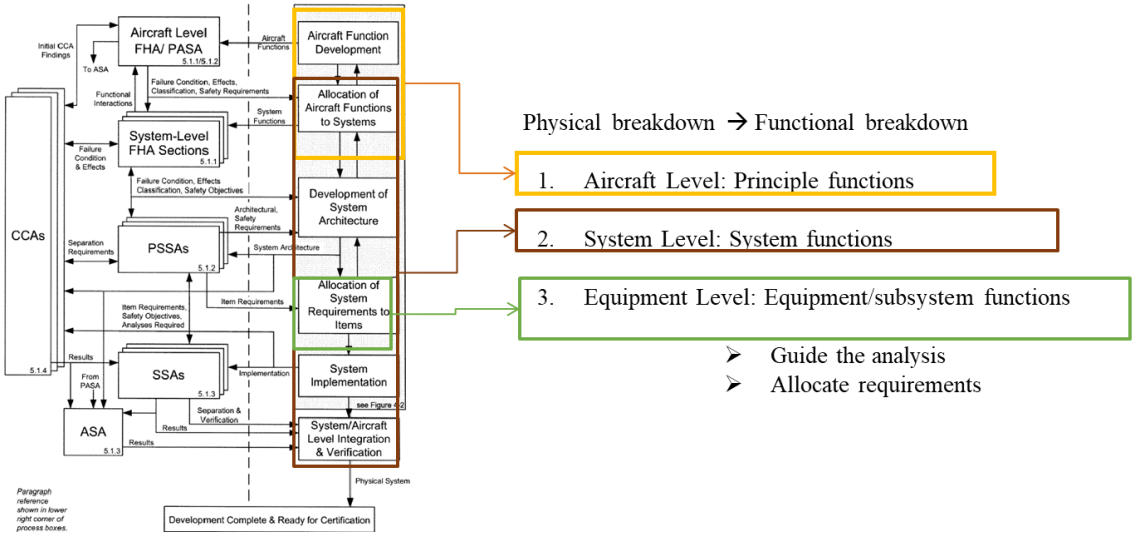

Figure 3: Global safety process in Aeronautics

2.2.2.1. **Functional Hazard Analysis (FHA)**

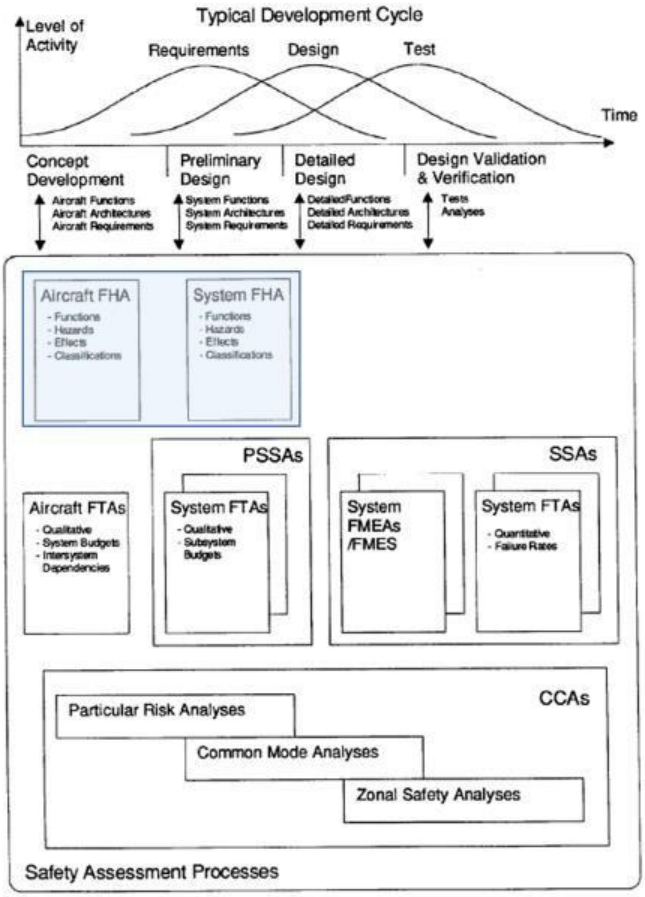

Figure 4: FHA in the Safety Assessment Process

FHA is the first step of Safety analysis. Following these steps it is possible to define safety targets to meet for each function.

- Identify all functions;
- List failure Modes;
- Consequences of each failure mode;
- Associate a criticality level to each effect;
- Allocate mitigations to reduce the criticality to an acceptable level (Table 3);
- Tests can be identified according the criticality levels (simulation, flight tests, etc.);
- Methods are also identified to verify the compliance to the requirements (Fault Tree Analysis).

Probability (Quantitative)		Per flight hour 1.0	1.0E-3	1.0E-5	1.0E-7	1.0E-9
Probability (Descriptive)	FAA	Probable		Improbable		Extremely Improbable
	JAA	Frequent	Reasonably Probable	Remote	Extremely Remote	Extremely Improbable
Failure Condition Severity Classification	FAA	Minor		Major	Severe Major	Catastrophic
	JAA	Minor		Major	Hazardous	Catastrophic
Failure Condition Effect	FAA & JAA	- slight reduction in safety margins - slight increase in crew workload - some inconvenience to occupants		- significant reduction in safety margins or functional capabilities - significant increase in crew workload or in conditions impairing crew efficiency - some discomfort to occupants	- large reduction in safety margins or functional capabilities - higher workload or physical distress such that the crew could not be relied upon to perform tasks accurately or completely - adverse effects upon occupants	- all failure conditions which prevent continued safe flight and landing
Development Assurance Level	ARP 4754	Level D		Level C	Level B	Level A

Note: A "No Safety Effect" Development Assurance Level E exists which may span any probability range.

Table 3: Failure condition severity as related to probability objectives and assurance levels

2.2.2.2. **System Safety Assessment (SSAs)**

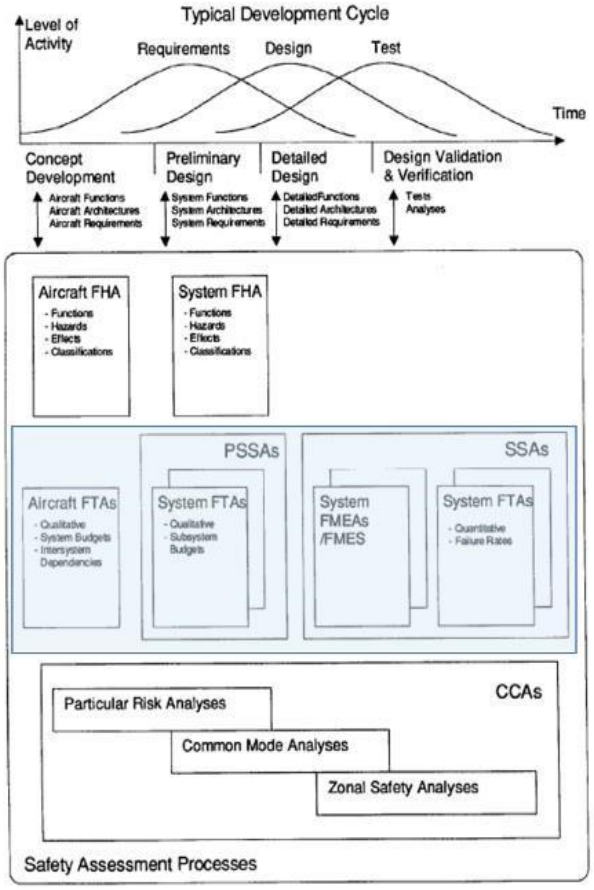

Figure 5: SSA and PSSA in the Safety Assessment Process

At this phases architectures diagrams become available. The SSA aim is to verify that all these architectures comply with safety requirements defined in FHA. This Preliminary Safety Assessment consist of:

1. Completing the list of Failure Modes and Safety requirements;
2. Listing pilot's action and maintenance to be carried out if a failure happens;
3. Verifying that system's architecture is in coherence with considered failure mode;
4. Allocating safety requirements at equipment level that will be communicated to the suppliers.

While System Safety Assessment consists in integrating suppliers' safety results to PSSA results to verify the compliance of the architecture with regards to the safety requirements (both quantitative and qualitative) defined in FHA and PSSA.

2.2.2.3. **Common Cause Analysis (CCA)**

Safety requirements may sometimes require independence among multiple function. The aim of Common Cause Analysis is to ensure independence or that the associated risk is acceptable. In particular, CCA allows to go through a checklist of external events that could cause a catastrophic or hazardous/severe-major failure condition. These events should be mitigated in order to meet Hazardous/Major failure conditions assigned budget. CCA is subdivided into three types of analysis.

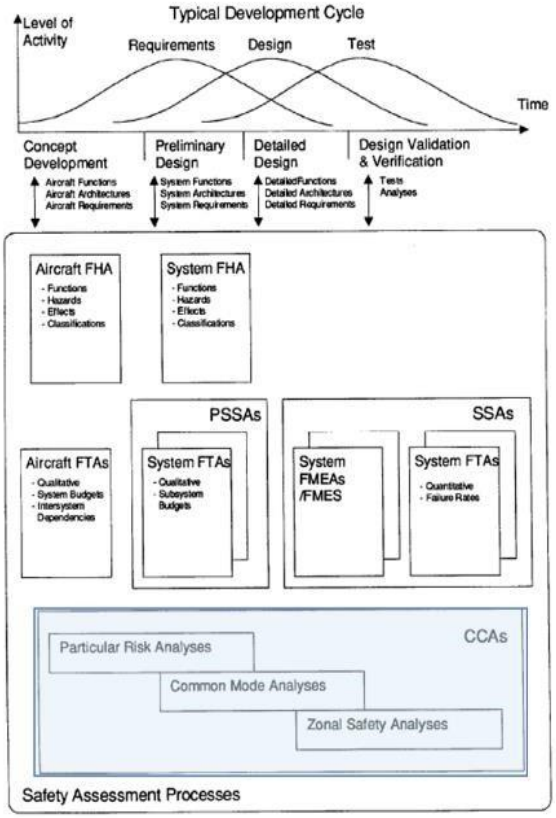

Figure 6: Common Cause Analysis in Safety System

a) Zonal Safety Analysis

Ensure that the installation of equipment, possible physical interferences with systems and possible maintenance errors are compatible with the independence requirements of the system.

b) Particular Risks Analysis

Verify that an external event (outside the aircraft or equipment) must not have any influence on the independence requirements.

Typical risks include, but are not limited to the following.

a. Fire
b. High Energy Devices
c. Leaking Fluids
d. Hail, Ice, Snow
e. Bird Strike
f. Tread Separation from Tire
g. Wheel Rim Release
h. Lightning
i. High Intensity Radiated Fields
j. Flailing Shafts

After identification of applicable risk; each risk should be subject to a dedicated study to examine the simultaneous or cascading consequences. The aim is to eliminate safety related effects or reach to an acceptable residual risk.

c) Common Mode Analysis

Verify independence of failures: equipment that are physically distinct but are composed of same hardware or software.

Here is a list of some common mode failures.

1. Hardware Error
2. Software Error
3. Hardware Failure
4. Production/Repair Flaw
5. Situation Related Stress (e.g., abnormal flight conditions or abnormal system
6. configurations)
7. Installation Error
8. Requirements Error
9. Environmental Factors (e.g., temperature, vibration, humidity, etc.)
10. Cascading Faults
11. Common External Source Faults

2.2.3. Methods and tools used for safety analysis

The safety assessment techniques can be classical as dependence diagrams or Fault Trees.
NB: Tools that can be used: Reliability Workbench

The safety assessment can also be based on models used to represent the functional and dysfunctional behavior of system.
NB: Tools that can be used: SIMFIANEO

2.3. Nuclear

Nuclear installation presents a specific risk as they all shelter more or less important quantities of radioactive substances that can provoke exposition of the workers, population or the environment to ionizing radiations and their effects.

2.3.1. Rules and Regulations

In France, the Law n°2006-686 of 13 June 2006 relative to transparence and the nuclear safety defines a complete legal framework, appointing the Autorité de Sûreté Nucléaire (ASN) to become an independent administrative authority. It provides regulations related to each phase of life, controls and also sanctions the Basic Nuclear Installation (INB).

The principal actors of nuclear safety are:

- The operators of nuclear installations, responsible of safety of their installations;
- The safety authority, whether it is civil of military, and it's experts commissions;
- The Institute of Radioprotection and Nuclear Safety (IRNS) expertise organisms;
- Local Information Commissions (LIC);
- The High Committee for the transparence and Information of Nuclear Safety (HCTINS).

2.3.2. Safety analysis process

The global process of prevention of accidents is based of defense in depth process which allows to compensate potential human and technical failures. It is based on multi-level protection centered where successive barriers block the radioactive substance leakage into the environment. Thus, one single failure cannot lead to an accident.

The Key Safety Function (KSF), if they are lost or degraded, can cause significant radioactive consequences to the environment. Therefore, there are multiple defense lines to avoid an Undesired Event (UE). These defense lines follow the principle of defense in depth process.

There are five important functions for safety to avoid an undesired event:

1. Guaranteeing the confinement of radioactive materials;
2. Guaranteeing the protection of external environment from ionizing radiations;
3. Guaranteeing the shutdown with sufficient sub-criticality and sufficient reactivity;
4. Guaranteeing the evacuation of thermal power;
5. Guaranteeing the prevention against the release of inflammable and toxic gases.

These functions drive the Elements Important for Protection (EIP). They can be structures, equipment, systems, materials, components and software that participate in a key safety function.
Actions Important for Protection (AIP) are all the technical and organizational measures of protection of interests.

In defense in depth each layer of protection is considered vulnerable, hence it is protected by another layer. As a result, the probability of that an undesired event will happen are extremely low.

There are 5 levels of defense:

- 1rst Level: Design and organization
  The first level consists in designing and building the plant with reliable techniques and materials and organize it operation in such a manner that it maintains the installation in its operational domain.
- 2nd Level: Control and Protection system
  The second barrier consists in maintaining the installation in its operational domain. It is based on controls, monitoring and protection to stop an abnormal evolution beforehand that materials are over constrained.
- 3rd Level: Backup system and accidental behavior procedure
  The third level is activated when the first two fail. It consists in backup and accident containment procedures.
- 4th Level: Limitation of serious accidents
  Even if very improbable, if the three first barriers fail and the core melts, the fourth layer tends to limit the releases in the external environment. The actions are part of Internal Emergency Plan elaborated by the operator.
- 5th Level: Limitation of radiation consequences for populations
  This last level consists in limiting consequences of radioactive releases over the people. The maneuver of this action means that all previous ones failed. These actions are grouped in Emergency Response Plan[2] (ERP).

2.3.3. Methods and tools used for safety analysis

The analysis methods and safety methods can be:

- Determinist: a risk analysis that identifies all the risk that can happen whatever is the cause of their surge - internal or external;
- Probabilistic: the evaluation is bases on a systematic investigation of accidental scenario.

The safety analysis takes in account technical dysfunctions but also Human Factor which designates human's behavior in their given duties at work.

2.4. Offshore

To anticipate dangers, humans rely more and more on simulations. These simulations are based on Hazop (HAZard and OPerability study), a methodology that consists in simulating all the accident risk and putting in place mitigation means.

2.4.1. Rules and Regulations

Following a serious offshore accident, European Union adopted the directive 2013/30/UE related to the safety of Oil & Gas operations in sea by modifying the directive 2004/35/CE. The aim of the directive is to prevent the serious accidents and limit their consequences by establishing the minimum safety requirements that can contribute indirectly the environmental and health conditions of workers during their duties in the sea.

[2] PPI: "Plan Particulier d'Intervention" in French

2.4.2. Safety analysis process

For offshore, risk analysis methods can be:

- Global: methods HAZOP and HAZID;
- Integrated: methods ARAMIS, LOPA, MOSAR, QRA.

2.4.2.1. Hazard and operability analysis (HAZOP)

HAZOP is a structured and systematic technique for risk management. It is best suited to thermos-hydraulic system for which it is crucial to control pressure, temperature, flow, etc.

It is similar to FMEA but considers probable deviation of major parameters important for the use of installation rather than failure modes. It focuses on expected conduct of procedures instead of expected function of equipment. Each system is break-down in several sub-systems called nodules. Then with the use of "guide words", parameters are modified out of the defined operation range. Deviation study aims to identify situations leading to potential risk for persons, goods and environment.

It is articulated around a HAZOP expert and Subject Matter Experts (SMEs) to predict deviations based on past experience and general subject matter expertise.

HAZOP methodology consists of four steps as described in the figure below:

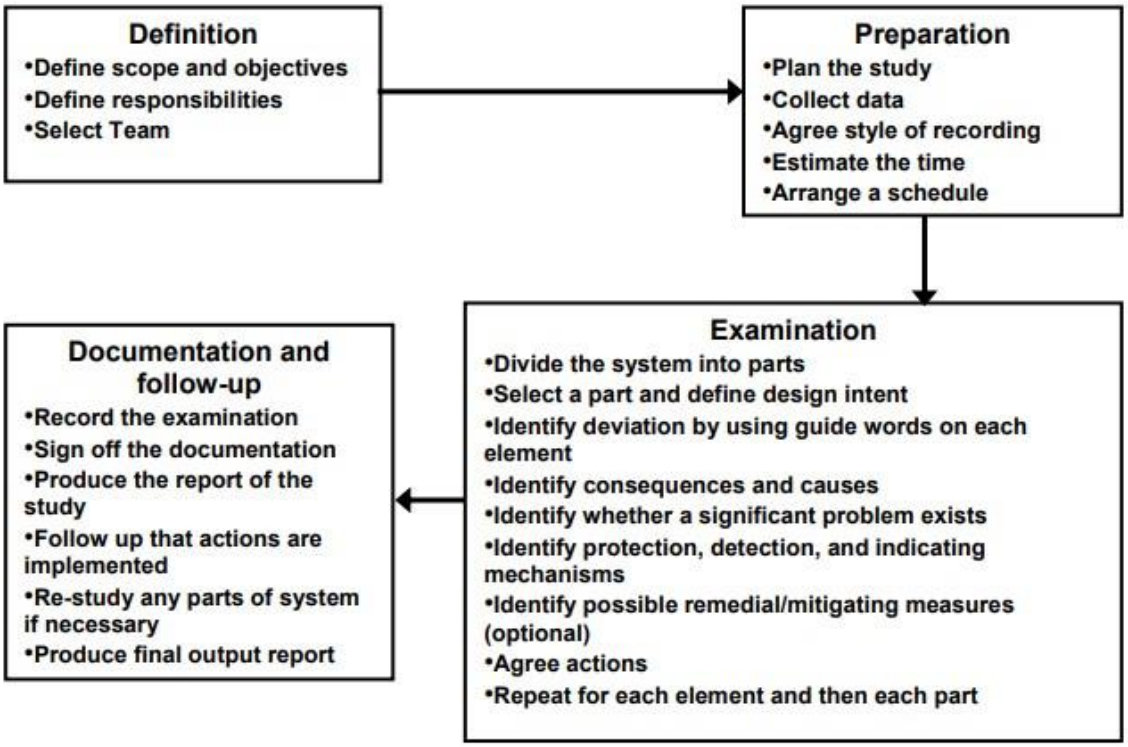

Figure 7: HAZOP methodology (from PQRI Risk Management Training Guides[3])

1. **Definition Phase:** this phase starts with preliminary identification of risk assessment team members. The team is multidisciplinary with SMEs that are the most knowledgeable hence relevant to predict the deviation. As it is a method conducted during a meeting, the climate should allow positive thinking and an open discussion.
   The team will then identify the scope of the assessment carefully and define boundaries and key interfaces as well as the assumptions.
2. **Preparation Phase:** The second phase consists in identifying the documentation to support the study, the user of the assessment results, project management, defining format to trace the study outputs and finally a consensus on HAZOP guide words to be used during the study.
   These guide words are a key element of HAZOP analysis, which according to IEC Standard 61882 "stimulate imaginative thinking, to focus the study and elicit ideas and discussion."

[3] PQRI Manufacturing Technology Committee – Risk Management Working Group Risk, Management Training Guides, HAZOP

Some common HAZOP guide words are:

Type of deviation	Guide word	Signification
Negative	No or not	Any part of intention is not accomplished
Qualitative modification	As well as	Simultaneous execution of another operation
	Part of	Part of intention is accomplished
Quantitative modification	More	Quantitative increase
	Less	Quantitative increase
Time	Early	An event occurs earlier intended time
	Late	An event occurs later intended time
Substitution	Other than	The result is different from the initial intention
	Reverse	Applicable to reverse of flow in a circuit or a chemical reaction

Table 4 : HAZOP example of guide words

NB: Others can be crafted as needed. Once the choice of guide words is final, the next phase can begin.

3. **Examination Phase:** this phase consists in identifying all elements (parts or steps) of the system or process understudy. A physical system breakdown leads to sub-systems. Processes breakdown leads to a number of separate steps. Similar steps and subsystems can be regrouped. Each step/part is then combined by guide word deviations in a systematic way. All scenarios do not necessary lead to credible situations but all those can lead to a credible situation by use or misuse of a step or part, should be documented.

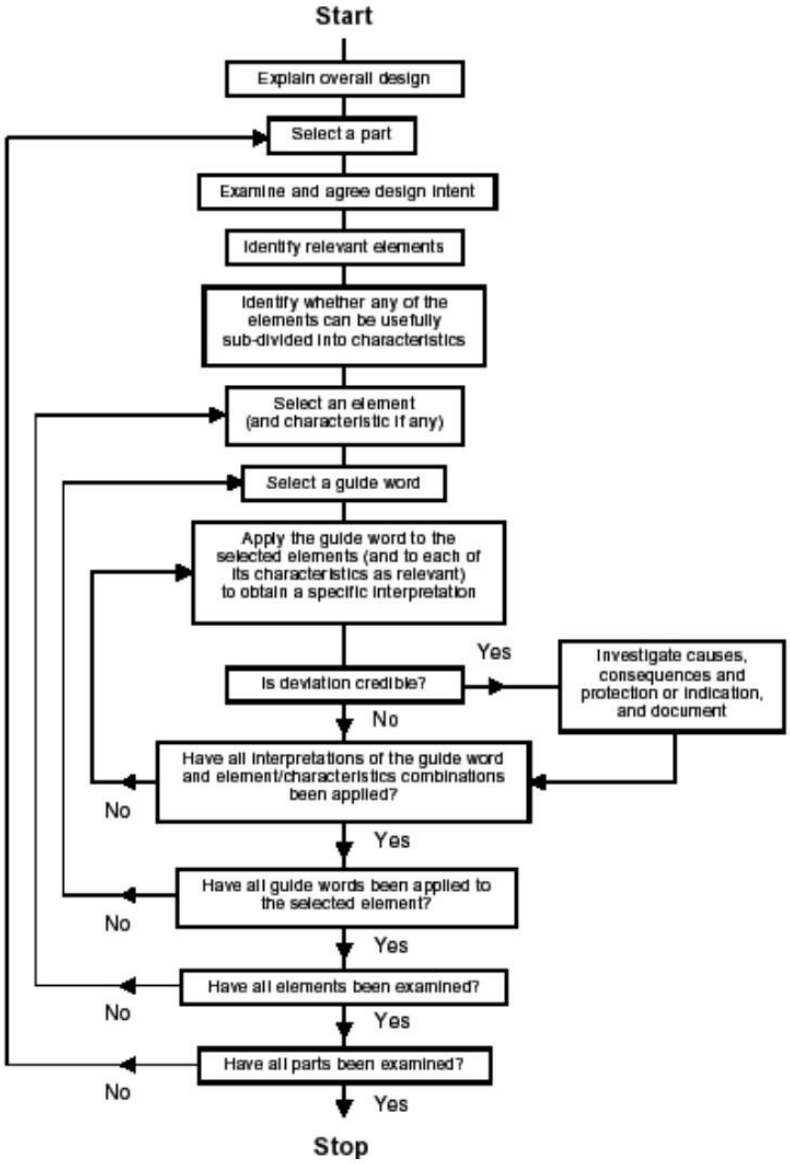

Figure 8: HAZOP examination phase process (from PQRI Risk Management Training Guides)

4. **Documentation and Follow up Phase:** The documentation can follow the templated given by IEC Standard 61882 and can be adjusted taking in account rules and regulations, need for more explicit risk prioritization, the company internal process, etc.

Once HAZOP method is completed, another process should ensure that the assigned actions are carried out.

2.4.2.2. **HAZID**

HAZID is a structured and analytical risk assessment method for identifying dangers. It aims to enhance:

- The impact of an installation on its environment;
- The impact of the environment on the installation;
- The interaction between the major elements;
- The general risks.

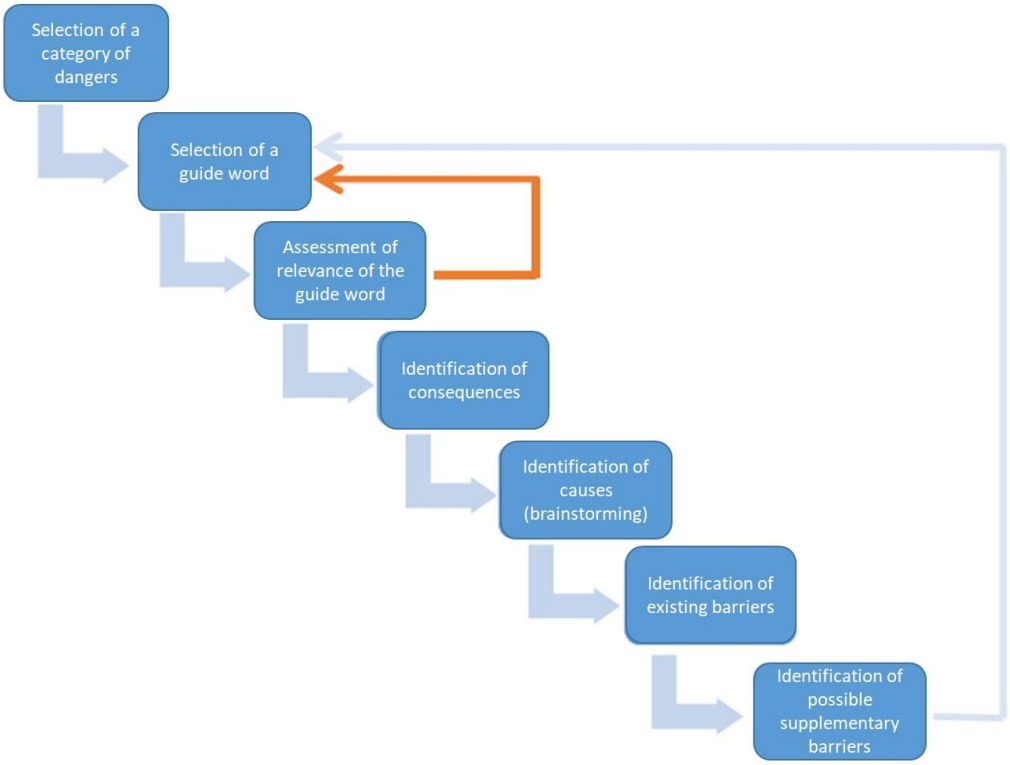

Figure 9: HAZID method

HAZID is similar to HAZOP in many ways but HAZID focused on external risks on the processes and the non-containment of the system. It evaluates consequences first then finds the causes of the incident.
It can be used in project phase as soon as enough information is available and even beforehand HAZOP to evaluate the implantation and certain installation choices.

2.4.2.3. **ARAMIS**

ARAMIS stands for Accidental Risk Assessment Methodology for Industries. It is divided in six main steps.

1. Identification of the major accident hazards (MIMAH)
2. Definition and evaluation of safety systems
3. Assessment of the management efficiency
4. Definition of the reference accident scenarios (MIRAS)
5. Risk severity mapping from the set of Reference Accident Scenarios
6. Vulnerability mapping of the plant's surroundings

A last step involves the crossing of this information for decision making.
The following figures explain further each step.

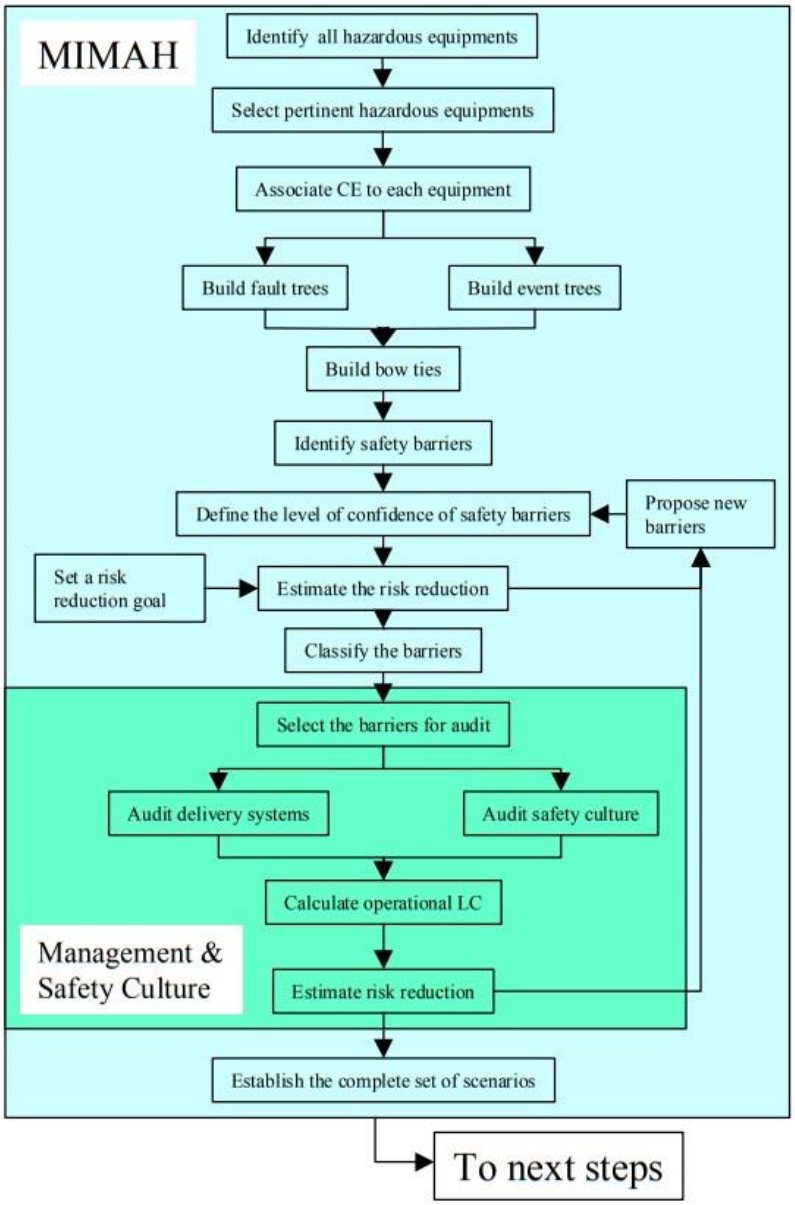

Figure 10: First steps of ARAMIS [4]

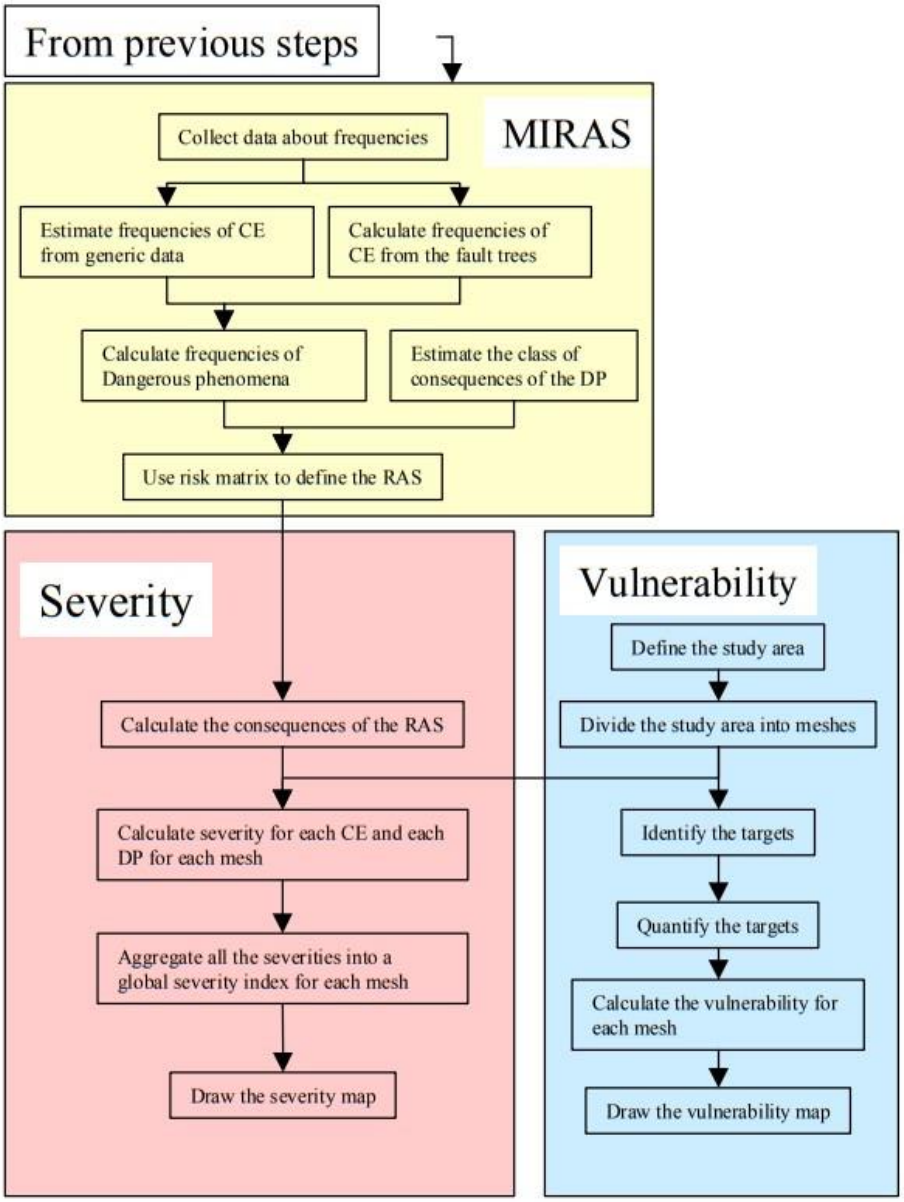

Figure 11: Last steps of ARAMIS [4]

2.4.2.4. **LOPA**

Layer Of Protection Analysis (LOPA) is safety barrier-oriented risk method but doesn't require a hazard criticality mapping.

The first step is to define selecting criteria for selection of scenarios to evaluate. These criteria can be a quantity of product (to be leaked), flow measurement or a consequence criterion that integrated within itself the information about the surrounding. The scenarios are then developed following traditional methods such as HAZOP or FMEA. A probability of occurrence of each cause is estimated according to experience, internal or external libraries. Identification of mitigation barriers follows this step. These barriers must be independent from the phenomenon or event happening. Their deployment and monitoring must be possible. The mitigation barriers that comply with these criteria are qualifies IPL (Independent Protection Layer). A failure layer is associated to them which corresponds to risk reduction factor. It is applied to accident probability. As a result, the SIL of the event is lowered of several levels. Matrices can help in

determining the minimum required SIL. The last step consists in ensuring that the risk is within acceptable criteria that have been fixed in the beginning. LOPA proposes four categories of criteria:

- A grid of criticality with acceptable limits in terms of frequency and severity;
- A quantitative criterion with regards to consequences of the scenario;
- A criterion specifying number of independent safety barriers to consider that the scenario is contained;
- A criterion of maximum cumulated risk for a site or process.

2.4.2.5. **Organized and Systematic Risk Assessment Method (MOSAR)**

MOSAR[4] allows to have a progressive risk assessment. It uses in an organized manner traditional techniques like PRA, FMES, and Fault trees and suggests grids and guiding lists. This method is composed of two modules that can be used more or less independently.

- 1rst module: Macroscopic analysis

It is a macroscopic risk assessment and is similar to preliminary risk assessment.

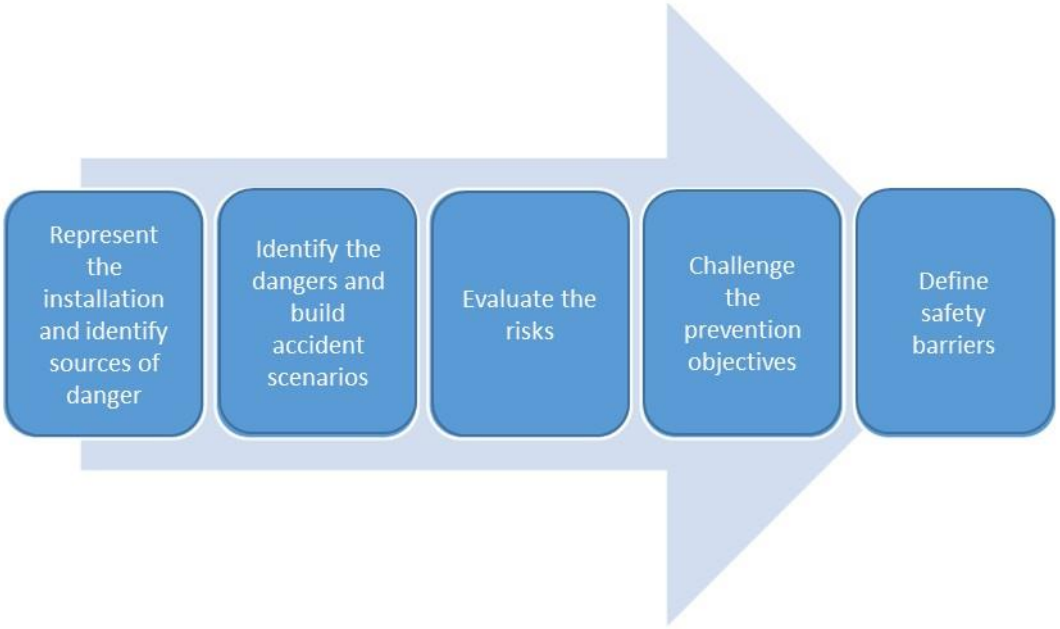

Figure 12: MOSAR method, macroscopic analysis

- 2nd module: Microscopic analysis

This is a more detailed analysis of scenarios identified in the first module with help of risk assessment techniques.

[4] Méthode Organisée Systématique d'Analyse de Risques in french

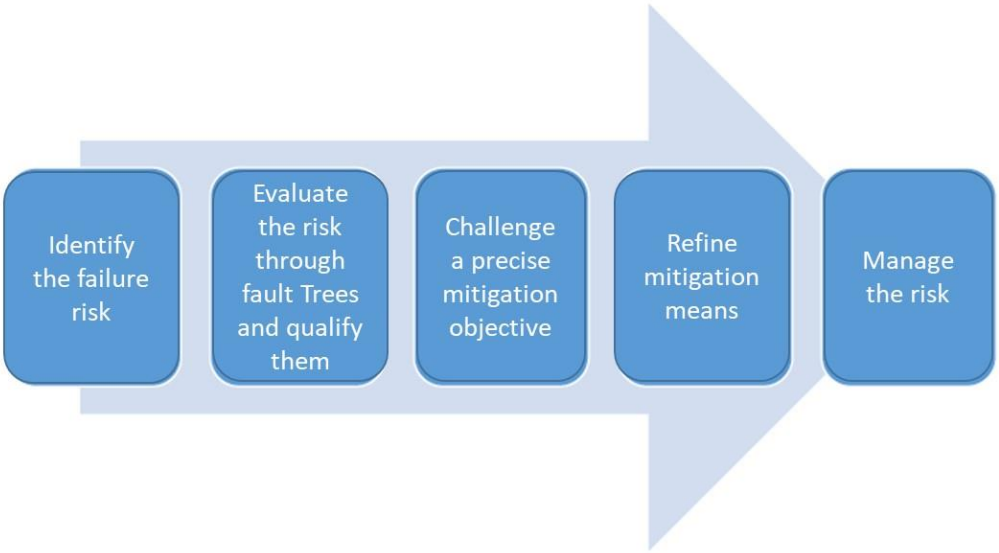

Figure 13: MOSAR method, microscopic analysis

2.4.2.6. **QRA**

QRA stands for Quantitative Risk Assessment and aims to determine the probability of damages caused by a potential accident. In this method, on the one side there are individual risks (probability of a fatality a given time and space from the consequences of an accidents) and on the other hand societal risk (numerous fatalities are caused by the accident).

Each installation is associated with an indicator taking into account the amount of storage or manipulation of dangerous substances, the type of equipment, and its exposition to any particular conditions, the state and nature of the substance.

For each selected equipment, a list of loss of containment events is elaborated. A probability is assigned to each event according to past experience.

For each event, the intensity of phenomenon is calculated through simulation of consequences. The intensity is expressed by distribution of toxic substance concentration, per level of thermal flows or by overpressure depending on the considered phenomenon. Simulation examples can be found in literature. Simulations are realized for different weather conditions associated with probability rates. Each set of Initial conditions is also qualified in terms of probability. The results are used to calculate the individual risk and societal risk.

The intensities of dangerous phenomenon are converted into probability of a single fatality or fatality of a fraction of society.

The last step estimates individual risk by summing up each probability of a single fatality risk. The societal risk determined by dividing the space around the installation in cells of same surface and evaluating the population exposed and the number of fatalities in each cell.
The sum of fatalities in each cell for all the considered conditions for a given scenario leads to the fatality toll for this set of conditions. For the societal risk, fatality classes are established (1, 10, 100 and 1000 fatalities) and sum of scenarios probabilities that could lead to at least as much fatalities of the class.

The results can be presented either in shape of a mapping of individual risk or by a graph occurrence rate/ numbers of fatalities. In both cases, the use of these results implies defining a mortality probability threshold for risk acceptance.

2.4.3. Methods and tools used for safety analysis

Safety barriers are identified through modelling and simulation to mitigate a disaster. It is possible to evaluate the propagation of an undesired event, the associated effect radius in order to decide the safety barrier that is the most relevant to be implemented.

A model can be digital or experiment based.
NB: Tools that can be used: ADMS, Phast, Effects…

2.5. Robotics

The safety within robotics aims to prevent damages of the robot itself but also all its surrounding environment and people around it. For robots used for servicing or medical assistance the consideration of interaction human/robots is necessary. This complexity tends to increase with the robot's capacity to handle more and more tasks.

2.5.1. Robotics as assistant to humans

2.5.1.1. Rules and Regulations

Since 2014, the international standard ISO 13482 describes dangerous phenomena and the way to deal with them for each type of assistant robots.

2.5.1.2. Safety analysis process

The global safety analysis consists in 3 major steps:

a) The appreciation of risk

Identification of dangerous phenomena and estimation of risk must be realized at each design phase and should consider:

- Uncertainty of autonomous decisions taken by the robot and dangerous phenomenon caused by the wrong decisions;
- The user and other exposed people knowledge, experiences and physical states;
- A spurious (not commanded) movement of the robot;
- An unsafe movement close to robot by people or animals;
- Adverse moving surfaces or conditions (for moving robots);
- The compliance to the human anatomy and its variations;

An estimation of risk should be associated to the above dangerous phenomenon. Mitigation barriers are deployed to reduce the risk so that the residual risk reaches an acceptable level.

Mitigation means	Guide word
Intrinsic Mitigation	The most important measures in the risk mitigation because they are likely to be effective permanently
Supplementary mitigation measures/ protection barrier	Robot commands protection function reducing significantly a particular type of risk.
User Information	Information about residual risk after application of intrinsic and protection barriers.

Table 5: Mitigation measures for safe operation of assistant robot

b) Validation of safety requirements, allocation of performance levels and implementation of mitigations; 2.5.1.2.3 Safety Performance

After the appreciation of risk, a Performance Level (PL) or Safety Integrity Level (SIL) must be determined for the following system functions:

- Emergency stop;
- Protection loss;
- Workspace limitations;
- Speed control related to safety;
- Force control related to safety;
- Avoiding dangerous collisions;
- Stability control.

c) Verification and validation of performances values related to robot's safety

All the performance values related to the safety of robot must be verified and validated by a process of risk reduction techniques corresponding best to the type of hazard.

Compliance to safety requirement can be verified by multiple methods:

- Inspection: Verify the stop function just by sensory control.
- Practical tests: the robot is tested in normal and abnormal conditions by injecting faults, by testing endurance, by testing breaking, etc.
- Measuring: compare the real values to the performance values.
- Observation during function: verify the state of robot just with the help of sensory control.
- Circuits' schematics review: examine the schematics and associated specifications.
- Software review: examine the design of code of software and associated specifications.
- Review of appreciation of risks with regards to the tasks: Examine risk assessment and the relevance of documentation.
- Review of implantation plans and relevant documents.

2.5.2. Programmable medical devices and electromedical systems

2.5.2.1. Rules and Regulations

Electromedical programmable systems follow a strict certification. ISO 14971 and EN 60601 that form the guidance and risk management method.

2.5.2.2. Safety analysis process

The supplier of medical robots must establish, document and maintain throughout the lifecycle of its product, a continuous process of risk assessment.

The global risk management process is patterned as the lifecycle of devices and consists of 5 parts:

- Hazard analysis[5]*;
- Hazard assessment*;
- Hazard mitigation;
- Evaluation of global residual risk;
- Production and postproduction information.

2.5.2.3. **Methods and tools used for safety analysis**

Usual safety management techniques can be used such as Preliminary Hazard Analysis, Fault Tree Analysis (FTA), Failure Mode, Effects Analysis (FMEA), HAZOP or Hazard Analysis and Critical Control Point (HACCP).
Model based analysis are also developed as HAZOP-UML based on the transcription of HAZOP guidelines in the context of different UML models.

3. Safety concerning the AI

This paragraph aims to focus ongoing safety efforts. In order to do this AI Act [5] is a corner stone as it is not specific to any industry, it addresses AI for all types of applications, but rather classifies AI by risk and type of AI. It has a risk based approach.

- The "unacceptable risks" are banned (Art 5 No.1 AI Act [5]).
- For high risk AI systems, mandatory requirements have to be met before its entry into service. (Art 18(1) AI Act [5]) and have a conformity assessment conducted by a third party or a notified body.
- Low risk AI systems do not have much restrictions.

It lays a framework for development, use and safety of trustworthy AI. It is equivalent to New Legalization Framework with single European Market [6].

The Annex 1 of the AI act [5] describes different techniques of AI and defines it as a "software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with" (Art. 3 No. 1 AI Act [5]).

3.1. General AI safety intended standards

This paragraph addresses the general safety intended standards that give us safety requirements or at least goals to achieve.

3.1.1. ISO/IEC TS 4213 [7]

This standard addresses AI based systems and specifies a methodology to classify performance of the machine learning models, systems and algorithms. It gives a framework on how to evaluate the machine learning of an AI via simple steps and simple advices.
It also gives a simple process to go through while assessing an AI as shown in the Figure 14.

[5] Analysis and assessment are different. To analyze is to perform a study of something to learn about its parts, what they do and how it they are related whereas an assessment estimates the value or character of the object. https://www.cp-journal.com/importance-analysis-versus-assessment/

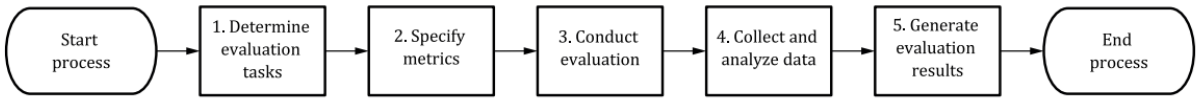

Figure 14: Generalized process for machine learning classification performance assessment (Figure 1 from the ISO/IEC TS 4213)

It also gives leads on how to measure performance through statistical measures of performance.

3.1.2. ISO/IEC 42001 [8]

This standard derivate from the ISO9001 [9] which gives a set of rules about quality management. This means it can give safety requirements to aim for a better quality of AI.
This standard gives a first normative draft on the implementation of particular process for the use of AI since it gives us guidelines on the use and the integration of IA-based systems in a product.
It answers some questions on the autonomy of AI-based systems and their transparency and explicability as well as the trust issues.
It is composed of 4 annexes:

- The annex A that lists objectives and check points for a trusty AI
- The annex B that gives guidelines on implementation of the check points and on the management of the data applied to AI.
- Annex C that details all the most common risks sources and proposes organizational objectives helping on tackling them during the AI development
- Annex D which details the domains in which the ISO42001 can be applied and its links to other standards.

3.1.3. ISO/IEC 23053 [10]

This standard is currently evolving.
It aims to establish AI and ML framework to describe a generic AI system using machine learning technology.
This framework describes how the system and components should be as well as how their function work in the AI ecosystem through definition and theory.

3.2. Aeronautics

AI can be very useful for aviation, especially Machine Learning could be used for design and operation, production and maintenance, Air Traffic Management, drones, Urban Air Mobility and U-space, safety risk management and cybersecurity.

3.2.1. EASA safety

One of the industries that is very keen to all safety aspects is aeronautics. EASA (European Aeronautics Safety Agency) has taken this innovation very seriously. In 2018, EASA had set up an internal task force to create EASA AI roadmap as a starting point for discussing European strategy for AI among EU stakeholders. The implementation of AI in aviation will impact EASA's activities such as certification, rulemaking, organizations approvals and standardization. Hence, it is best to assess the impacts and build the roadmap with associated actions.

EASA published a first version of AI Roadmap [11]. It takes in account AI HILEG findings and sets 5 AI Roadmap objectives to develop further the regulatory framework. EASA has a human

centric approach and is involving industry and research stakeholders to join the AI regulatory work. This roadmap has been updated in 2023 as the AI Roadmap 2.0 [12].

Some actions of the action proposed in roadmap are listed below:

- To establish public trust in AI based systems
- To integrate ethical dimension of AI in certification process
- To build a certification system for AI solution
- To identify the standards and protocols, methods that will be needed to evaluate that AI solution will improve current safety analysis

The requirements are non-binding at this level but rather for all stakeholders including regulators to test and provide feedback to improve. The higher level of trustworthiness is achieved the higher will be the social acceptance level.

The following diagram summarizes the seven key guidelines (Accountability, Technical robustness and safety, oversight, Privacy and governance, non-discrimination and fairness, transparency and societal and environmental well-being) which will first interface with AI trustworthiness analysis. Which itself serve as gate to three technical building blocks namely Learning Assurance, AI Explainability and AI safety risk Mitigation. The Top 5 AI Roadmap objectives are also listed in the diagram below. EASA has a human centric approach and is involving industry and research stakeholders to join the AI regulatory work.

Figure 15 : EASA Roadmap and AI objectives

As a way forward, EASA considers that under the key field of "oversight" human machine relationships may evolve. Three scenarios are considered: Humans can be in the loop (HITL), human on the loop (HOTL), and Human in control (HIC). Regarding these relationships, we could classify the AI/ML application as in the table below ideally increasing the automation. It could apply to autonomous flight but also to other industries.

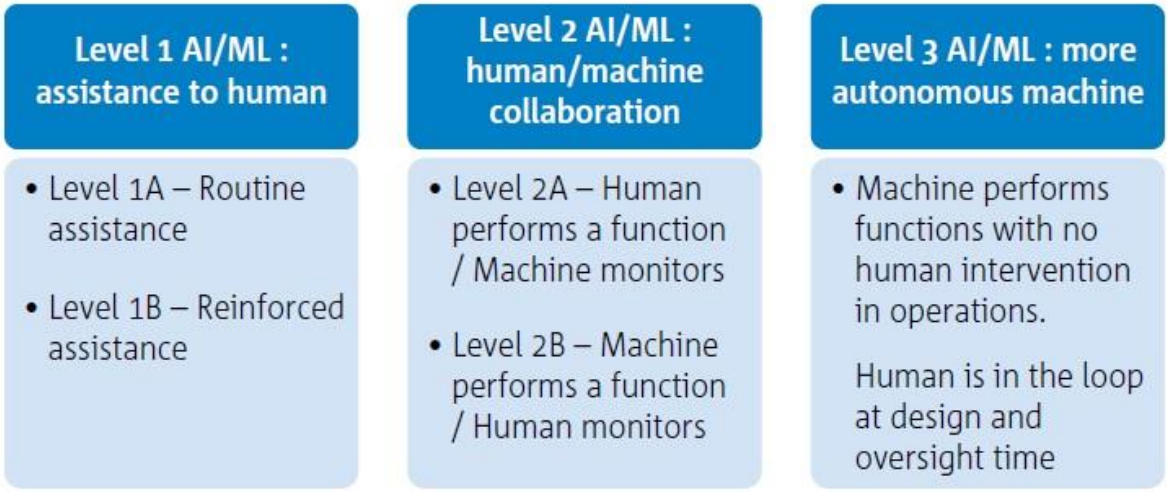

Figure 16:AI/ML classification suggested by EASA roadmap

Civil Aircraft certification pilot projects with limited AI/ML use are already launched. To reach the autonomous flight phase, most of them will go through the single Pilot Operation with support of virtual co-pilots (2025), then with autonomous flights with pilot supervision and finally with fully autonomous flight (2035). Concerning drones for autonomous flight, industrialists would like to reach it as soon as possible. The condition would be to keep a gradual strategy for CAT scenarios. For UTM field, an initial guidance has been issued in 2021 to support the first applications of U-space and automated/semi-automated drones. EASA has published after EASA AI Roadmap 1.0 three more focused deliverable on methods for meeting safety compliance.

1. EASA Concept Paper: First usable guidance for Level 1 machine learning applications [13]
2. Concepts of Design Assurance for Neural Networks (CoDANN) [14]
3. Concepts of Design Assurance for Neural Networks (CoDANN) II [15]

Other guiding material is expected to be published following this timeline in Figure 17.

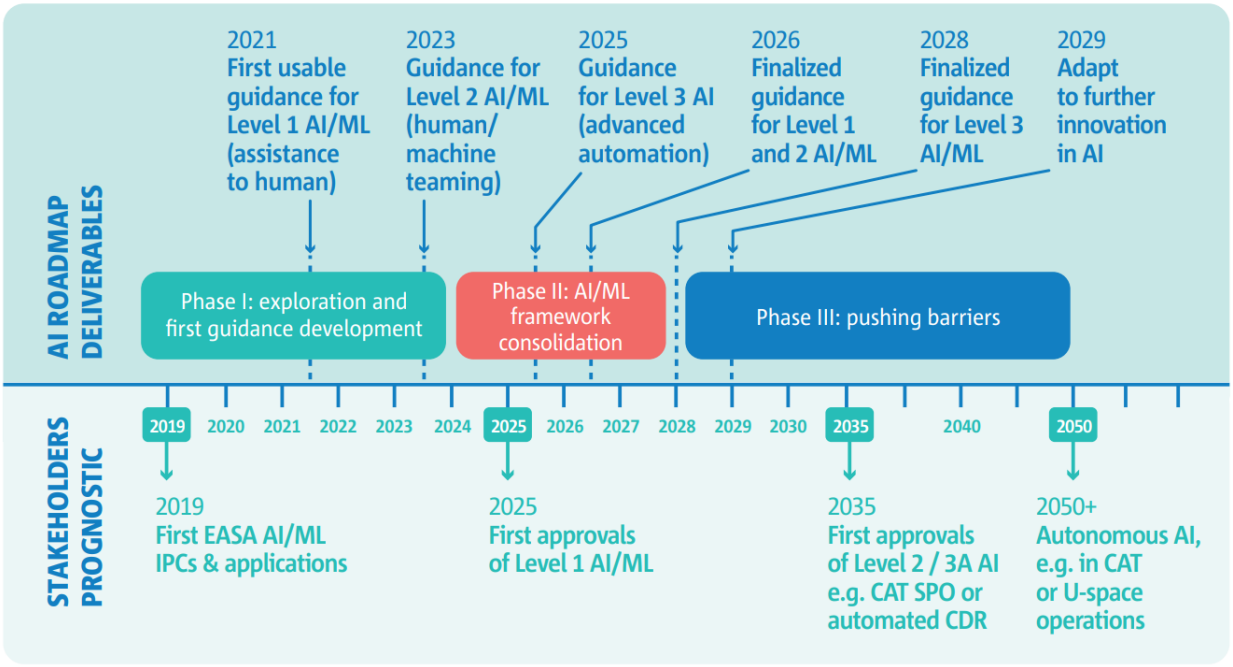

Figure 17: Timeline set by EASA AI roadmap

3.2.2. Formal Methods Use for Learning Assurance (ForMULA)

Together with the Collins Aerospace, the EASA published a report on formal methods use for learning assurance (ForMULA - [14]) on April 2023.

This report focuses on gathering the existing formal methods used to demonstrate the compliance for the assurance and certification objectives for machine learning.

It details:

- a state of the art of the formal methods specific to ML from validation and verification to machine learning stability and robustness verification and explainabilty, as well as quantitative methods to do so.
- A state of the art of the formal methods on learning process with a special regard on stability of the machine learning model stability
- Practical demonstration and use of formal method on a use case of deep learning.

3.2.3. Machine LEarning application APproval MLEAP

To follow the objective of the AI Roadmap a workgroup composed of the EASA and Airbus has been made to produce reports propose a way to certified or qualify an AI based system in the Aeronautic.

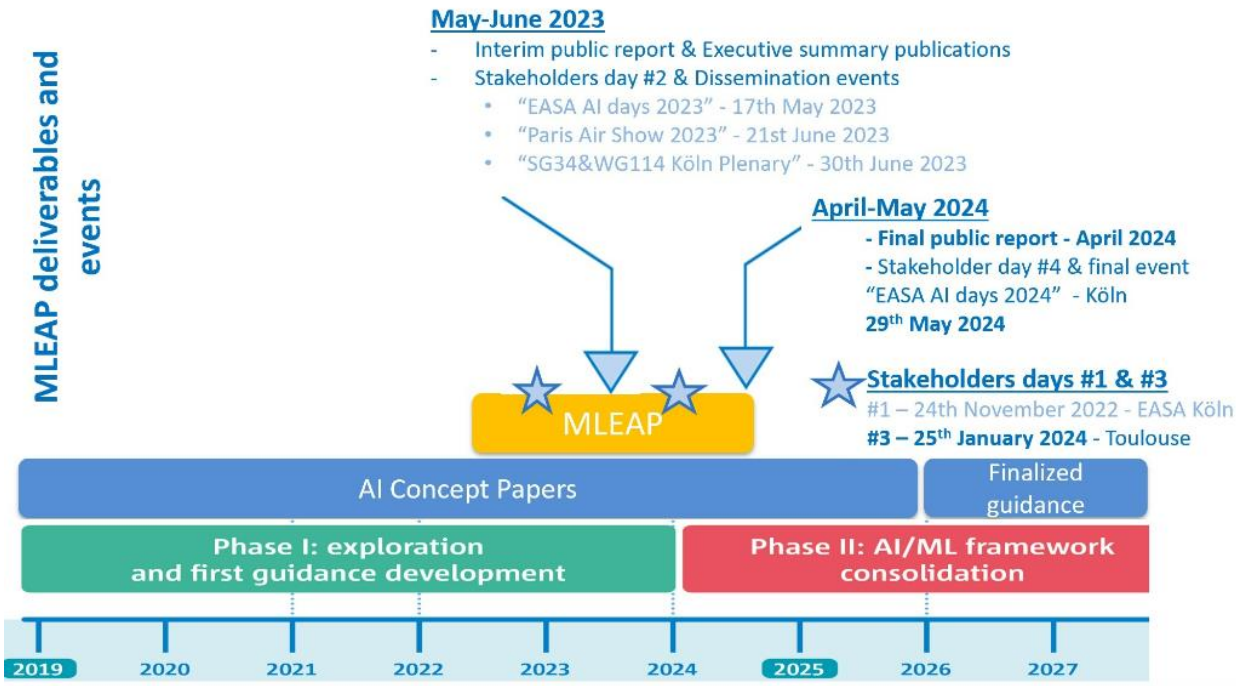

Figure 18: MLEAP in the AI roadmap from the EASA

This workforce called MLEAP has produced many reports and development that aim to regroup and assess state of art on AI based system certification on 3 axes:

- Data completeness and representativeness
- Generalization properties of the model development
- Robustness and stability on the AI evaluation process

3.2.4. Eurocontrol

The European Organisation for the Safety of Air Navigation commonly known as Eurocontrol is an international organization working to guarantee a safe and seamless air traffic management across Europe. They organize every year webinars or forums named FLY AI. They published in March 2020, The Fly AI Report - Demystifying and accelerating AI in Aviation/ATM [17] for the first time. Last version is from 2021. After dedicating chapters for use cases for AI and potential axes to work on this report discusses how to guarantee the safe use of AI. They believe that AI can improve safety joining EASA conclusions. Given that not only it can process huge amount of collected data but also can be fed with historic data to identify patterns and also weak signals.

Eurocontrol proposes a process specific for AI-based products: their recommendation for R&I are to first focus on AI potential in safety critical operations and secondly to focus on new methods to prove for the validation and certification of advanced AI applications to ensure their transparency, robustness and stability under all circumstances and development of Safety Intelligent tools.

In terms of validation and standards, it joins EASA actions to develop guidelines that encompass the full scope of aviation. More specifically to adapt current Air Traffic Management safety cases to AI based solutions. This will require short and quick loops to provision frequent updates of AI algorithms, development of validation and verification tools for such algorithms including

and probabilistic analysis, bridging the gap between the industrialists and the regulators, providing a legal framework for AI companies.

In terms of automation levels this report breaks it down to 6 levels numbered from 0 to 5. Which is the double of what EASA suggests. In the first three levels actions can only be initiated by Humans.

3.2.5. The European Organisation for Civil Aviation Equipment (EUROCAE)

The European aviation industry stake holder and normative giant SAE have recently launched EUROCAE WG 114/SAE G34 to come up with a process for certification and/or approval of AI-based products. This could include learning assurance, formal methods, testing, explanation, licensing, in-service experience, and online learning assurance. Depending on the product and the inputs from the trustworthiness analysis only relevant activities may be carried out and the demonstration level could also be adjusted. The figure below lays out the process:

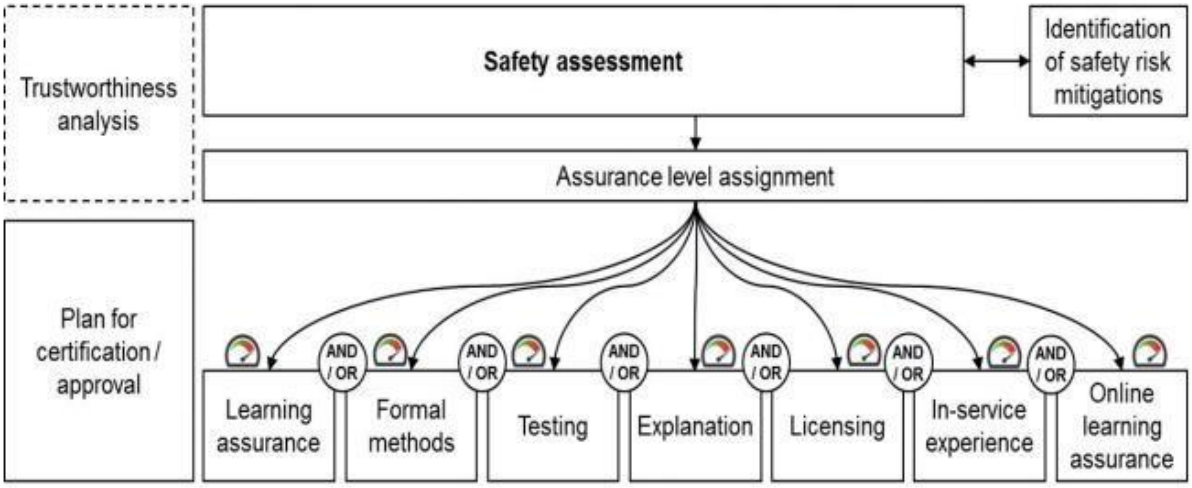

Figure 19 : Fly AI proposal for the future process for AI based products

3.2.6. SAE ARP6983 [18]

This document is a standard which is a work in progress.

It is a standard that will propose guidelines for the development of Aircraft Systems leveraging AI capabilities, while taking into account the overall aircraft operating environment and functions. It will give guidelines on validation of requirements and verification of the design implementation for certification and product assurance and guidelines with the assessment of safety.

It will show examples on how to show compliance with the regulations and will assist companies in developing and meeting its own internal standards by considering nowadays guidelines

3.3. Rail

The Rail sector could benefit from the AI technology for enhancing the capacity, speed and safety. Among many applications we can quote an early obstacle detection and accurate position determination [6]. These functions assembled with other could allow the train to run driverless across Europe.

Similarly, to autonomous vehicles autonomy levels there are Grades of Automation for trains. The grading goes from 0 (no automation) to 4 (full automation). It starts with the level GoA 0, where the driver is in command of the movement of train and its supervision. GoA 1 describes a manual control with a separation and speed supervised by the automatic systems. At GoA 2, we reach a semi-automatic train when the train can ensure its traction and braking. But the driver is in charge of supervision and can take control of the train. At GoA 3, the train takes over the responsibility of monitoring its environment and detect and react to obstacles in an autonomous level. There no more driver but personal for doors control and emergency operation. Finally, GoA 4 is the stage of full automation. There's no need for on board staff. The train performs all the tasks by itself. The Tableau 1 describes the relation between automation grades and the functions of a train.

For the moment EU 2016/797 application for the regulation of safety component for high risk AI seems is excluded in the section B Annex II of AI Act [5]. So elements that could fall under this regulation have to find another EU safety product regulation authorized by AI Act [5] to be compliant.

Conventional railway-specific safety standards (EN 126, EN 50128 and EN 50129) do not specify neither any AI system definition in terms of techniques nor any delimitation perimeter of functions or conditions to apply. EN50128 does take the eventuality to use AI but do not specify any special conditions linked to its use. Thus, the AI application currently have to follow the same standards and comply with the same safety requirements. The definition and associated demonstrations for Tolerable Hazard Rates described in 2.1.2.4 remain the same.

In conventional process all components have to demonstrate that they fulfill system requirement specifications and safety requirement specifications. AI has to comply with these demonstrations and be just as safe and transparent as other systems. The challenge is that where determining input and outputs can be demonstrated, the processing and tracing of an AI decision based on Convolutional Neural Network is not yet possible. That's why, AI is sometimes designated as a black box.

Basic functions of train operation		GoA0	GoA1	GoA2	GoA3	GoA4
		On-sight train operation	*Non-automated train operation*	*Semi-automated train operation*	*Driverless train operation*	*Unattended train operation*
Ensuring safe movement of trains	*Ensure safe route*	X (points command/control in system)	system	system	system	system
	Ensure safe separation of trains	X	system	system	system	system
	Ensure safe speed	X	X (partly supervised by system)	system	system	system
Driving	*Control acceleration and braking*	X	X	system	system	system
Supervising guideway	*Prevent collision with obstacles*	X	X	X	system	system
	Prevent collision with persons on track	X	X	X	system	system
Supervising passenger transfer	*Control passengers' doors*	X	X	X	X	system
	Prevent person injuries between cars or between platform and train	X	X	X	X	system
	Ensure safe starting conditions	X	X	X	X	system
Operating a train	*Set in/ set off operation*	X	X	X	X	system
	Supervise the status of the train	X	X	X	X	system
Ensuring detection and management of emergency situations	*Perform train diagnostic*	X	X	X	X	System and/or staff in OCC

Table 6: Train operation for the different grades of Automation [19] [6]

To compensate this lack in standards and support of its regulation, the European union has initiated partnerships. We can quote for the rail:

- Shift2rail: It was a partnership that ended in 2021 and followed European partnership on rail research and innovation established under the Horizon Europe program that will end in 2027. This partnership aimed to accelerate research and development in innovative technologies, including work on AI in the rail. It supported the fulfilment of European Union regulations and objectives relevant for the rail. It also supported the Single European Railway Area, by accelerating innovation on rail technologies.

[6] OCC: Operation Control Centre

- Europe Rail: it is the follow up of the project Shift2Rail, a new partnership with the same objectives.

Another project, FP1-Motional, that is focused on advancement in the rail technology and its operation to improve its efficiency, safety and sustainability, addresses high level requirements for design, as well as use cases to enhance the rail. The high-level specifications directly incorporate the state of the art and the analyses and incorporate the results of Shift2Rail.

3.4. Agriculture

With the ISO 18497 from 2018 the agriculture industry has prepared itself from tackling AI-safety in treating the safety of highly automatized systems.
This standard specifies design principles on the highly automated aspects of highly automated machines and vehicles. It gives a safety framework build around requirements, means of verification and information about the use to ensure a satisfactory level of safety on agricultural and forestry vehicles.

3.5. Medical

With the IEEE 2801 and the IEEE 2802 published in 2022 the medical world has set foot in the IA domain. These standards aim to standardize the quality management of activities for dataset used for AI in medical devices by defining objectives to attain for a good database but also the good use of the data.

By extension, they give a set of generic requirements for the data used for the machine learning in the medical industry.

3.6. Conclusions on safety the AI domain

If a solution for making deep learning AI decisions explainable and understandable has to come to existence, one thing is sure: it will take time. So, a mid- term solution should be found. The drivers may stay a little longer in the loop and validation & verification focus may shift from transparency to extensive testing and acceptable accurate response rate. Which is also a mitigation suggested by EASA. Meanwhile AI specific standards are developed, conventional standards may extend its requirements to be more inclusive of AI use.

Ongoing efforts for regulations are developed in deliverable 1.3.

All those industries have engaged in initiatives to identify impacts and challenges that AI bring to their applications.

In all those industries, among other things, AI could improve the safety. The timely obstacle detection and accuracy of data processing could lead to safer operation. However, we cannot guarantee that AI is safe in itself.

The current pace of convention standard making might be too long. Short-term and mid-term safety consensus should be found by the authorities and met by the industrialist.

4. Safety evaluation/Certification of AI

The previous chapter describes a state of art of hazard analysis in different domains. This chapter aims to present a mapping of certification and labelling initiatives for AI and will present the case study of LNE certification for AI evaluation processes.

4.1. Mapping of labelling and certification references

LNE already performed a mapping of labelling and certification references [20], that the deliverable aims to update.

AI offers great functionalities for robotics (industrial, medical, assistance) for Autonomous system (Urban air Mobility, Autonomous Vehicles). Their performances, however need to be demonstrated in terms of ethic, explainability, resilience. In this way, users can be granted acceptable guarantees regarding the products. This need is driving institution and industrial stakeholders to work on labelling and certification of such products and processes.

This mapping is based on several axis:

- What type of mark does it give:
  - Certification (Mandatory or optional)
  - Label (given upon complying to a list of specifications)
  - Other (when not enough information is available to classify)
- The segment covered by the certification or label:
  - The product containing IA developed by the supplier applying for certification;
  - The process implemented by the supplier before, during or after commercialization (design, operational conditions, etc.);
  - The Management system: means a particular type of organization, largely adopted in 27001, 50001 etc;
  - The company itself. At numerous occasion the process lead to granting the brand to the entire company, even if the verification concerned only one product or one particular process in the first place.
- The branding can concern different topics that can be inspected:
  - Performance
  - Safety
  - Security
  - Ethics
  - Transparence
  - Privacy
  - Quality of datasets
- Finally the research results are classified considering
  - The type of AI technology
  - The field of application

The table below aims to present a summary of this mapping.

Type	Name	Scale	Maturity	Target	Topics
Certification	Responsible Limits on Facial Recognition	Europe(Switzerland)	Started	Process	Performance, Security, ethics, transparency, Privacy, Quality of datasets
	Trusted Artificial Intelligence - Towards Certification of Machine Learning Applications	Europe(Austria)	Started	Product	Ethics, transparency, Privacy, Quality of Datasets
	Certification des outils et des services d'intelligence artificielle dans les systèmes judiciaires et leur environnement	Europe(European Commission)	Started	Product, Process, Management system	Privacy
	Requirements Management for AI assisted diagnostic technology	Asia(China)	Launched	Product	Performance, Safety, Security, Ethics, transparency, Privacy, Quality of dataset
	AI Certification standard for AI Processes (Référentiel de certification de processus pour l'IA)	Europe(France)	Launched	Process	Performance, Safety, Security, Ethics, Transparency, Privacy, Quality of dataset
	AI-ITA (AI Innovative Technology Arrangement)	Europe(Malta)	Launched	Product	Performance, Safety, Security, Ethics, Transparency, Privacy, Quality of dataset
	RAI Certification	North America(USA)	Started	Product	Security, ethics, Privacy, Quality of dataset
	360° Certification for Artificial Intelligence You Can Trust	Europe(Austria)	Started	Product, Process	Performance, Safety, Security, ethics, transparency, Privacy, Quality of dataset
	The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS)	International	Started	Product	Ethics, Transparency
Label	Label GEEIS-AI	Europe(France)	Launched	Management system	Ethics
	IA France	Europe(France)	Launched	Company	
	ADEL - Premium	Europe(France)	Launched	Process	Security, Ethics, Transparency, Privacy, Quality Da-tasets
	ISpaces	Europe	Launched	Product, Company	
	Label ScoreFact	Europe(France)	Launched	Company	
	Breakthrough Devices Designation	North America(USA)	Launched	Product	Performance
	Solar Impulse Efficient Solution Label	Europe(Switzerland)	Launched	Product	
	Ethics Label for AI	Europe(Germany)	Started	Product	Security, Ethics, Privacy
Other	IT-security	Europe(Denmark)	-	Process, Management system,	Security, Ethics, Transparency, Privacy

Type	Name	Scale	Maturity	Target	Topics
				Company	
	Vertrauenswürdiger Einsatz Von Künstlicher Intelligenz	Europe(Germany)	-	Process	Performance, Safety, Security, Ethics, Transparency, Privacy, Quality of dataset
	AI Act	Europe(European Commission)	-	Product, Process, Management system	Safety, Security, Ethics, Transparency, Privacy, Quality of dataset
	AI Readiness Index (AIRI)	Asia(Singapore)	Launched	Company	Quality of dataset
	Artificial Intelligence Ethics framework	Oceania(Australia)	-	Product	Performance, Safety, Security, Ethics, Transparency, Privacy
	Artificial Intelligence Ethics, governance and policy challenges	Europe(Belgium)	-	Product, Process	Performance, Safety, Security, Ethics, Transparency, Privacy
	AI Maturity Tool	Europe(Finland)	Launched	Company	Safety, Security, Ethics, transparency, Privacy, Quality of dataset

Table 7: Mapping of labelling and certification initiatives for AI

4.2. Focus on the possible processes of allocation of label or certificate

The allocation of a label or a certification can be conditioned to the verification modality for the compliance of the requirements described in the concerned standard.
The modalities can be self-declaration, literature review, audit or testing by the third-party assessor. The verification modality are defined at the time of design of the standard and described in the standard. The standard may also provide for a mixture of several means of verification (for example, audit and testing or literature review for certain requirements or audit for other requirements).

Below is a focus on the possible modalities in increasing order of the level of confidence provided by the verification modality:

a) Self-declaration: a requirement whose response mode is self-declared does not rely on any third party verification and leads to a zero confidence level in the result. Self-declaration has the advantage to leads the applicant to reflect on its level of compliance/maturity with respect to a requirement.
b) Literature review: This type of verification requires the applicant to provide compliant documentation and to follow its practices as described in its documentation. It has the advantage of being the easiest type of verification to implement by the verifier. Nevertheless, the level of confidence associated with this type of verification remains relatively low due to the simplicity of producing documentation whose sole purpose is to meet the labelling requirement and which is then not used or implemented by the applicant.
c) Audit: The audit consists of verifying the application of the requirements by the applicant by interviewing the people involved or taking evidence via internal tools (project management, development, etc.). It is also based on the review of internal documents. The methods implemented by the company are challenged by the auditor. The level of confidence that can be given to the audit is between medium and high depending on the duration of the audit and the skills of the audit team.
d) Testings: Testing is the final level of verification and confidence possible as the testing and technical evaluation of the developed product is carried out entirely by the verifier, thus providing maximum independence

In the case of the AI processes certification discussed in the next paragraph, the respect of all requirements by the certified organism is checked by means of an audit process.

4.3. Case study- LNE Certification standard of processes for AI[7]

LNE certification standard of processes for IA [21] describes how to certify AI the processes and define common requirements related to the design, development, evaluation and Maintenance in Operational Conditions (MOC) of AI systems based on machine learning algorithms in accordance with the needs of their users.
This standard covers all sectors of activity in which AI systems are used.
It is not intended to define requirements applicable to AI functionalities themselves and therefore specific to their uses.

[7]The LNE IA certification standard is available on the website of LNE:
https://www.lne.fr/en/service/certification/certification-processes-ai

The standard define as follows the processes covered by the requirements:

1. Design process: it consists in transforming an expression of need into functional specifications.
2. Development process: it consists of implementing these specifications into a version of the AI functionality ready for evaluation.
3. Evaluation process: it consists of verifying and validating the conformity of the system with the defined specifications before its deployment.
4. Maintenance process: it aims to ensure that the AI functionality complies with the defined specifications after its deployment and throughout its operational phase.

The applicable requirements are described and listed in the chapter III of the certification standard.
The requirements III.1 concern the definition of the processes included in the scope of Certification and statement of applicability. Through these requirements, the applicant shall document the processes covered by the certification application as well as their interfaces, the sites where they are applied, the personnel allocated to these processes, and provide a justification for any exclusion of applicability of the requirements of the standard. The applicant shall also implement and maintain the processes. To achieve this, the applicant must document the required inputs and expected outputs of the processes, determine the interfaces of the processes, identify the resources needed to ensure the proper functioning of these processes, evaluate the processes and implement all the actions required to ensure the capacity of its processes to achieve the expected results.
The requirements III.2 concern the requirements linked to planning, operation, evaluation of processes. In a global way, regarding the identified risks related to the use of the AI functionality, the applicant shall plan effective risk mitigation actions and the evaluation of the risk mitigation actions carried out.

This report focuses on the requirements for evaluation process but the reader is invited to consult the entire standard with regard to the other processes. In addition, for many of the requirements, the standard provides application examples that help with understanding/interpretation.

Regarding the requirements dedicated to the evaluation process and described in requirements III.5, the standard defines minimum inputs and outputs that are required to be documented by the applicant.
For input elements it shall provide:

- Functional specifications;
- Test data meeting the requirement
- A prototype with implementation AI function concerned by performance evaluation For output elements:
- Evaluation protocols, tools and metrics,
- Results of the evaluation process,
- AI function concerned by performance validation;
- The final risk analysis;
- The associated documentation (user manual, description of the models, etc.)

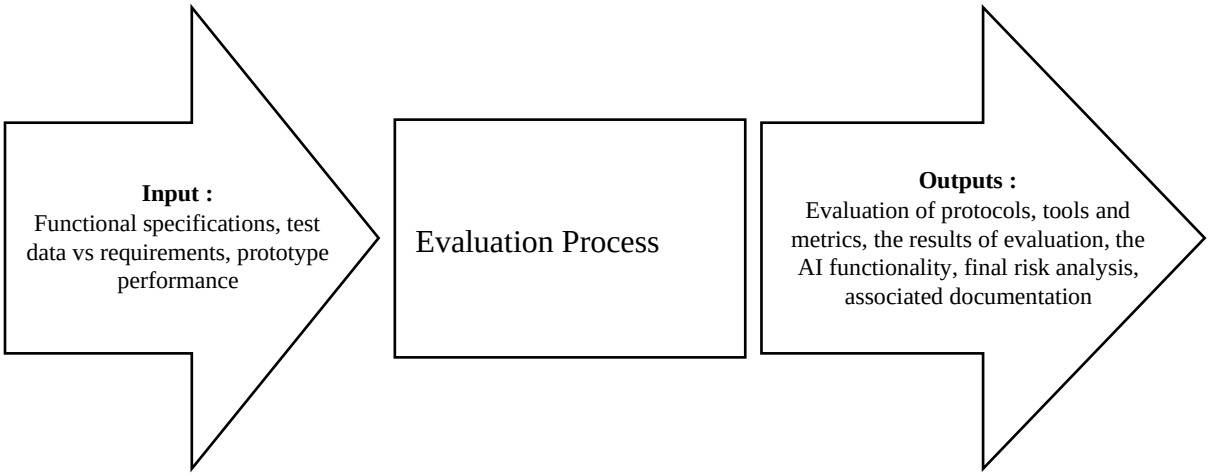

Figure 20: Evaluation Process inputs and outputs (LNE certification standard for AI processes)

The standard recommends a number of evaluation rules that an AI functionality supplier must follow. These requirements are presented in the table below. First column represents the requirements of evaluation to be observed by the certifying body and the second column describes the way to carry out these tasks.

Evaluation Requirements	Requirements for Applicant
Shall start after learning Process	Shall document the learning process (particularly for human interventions and performance verifications)
Shall use evaluation metrics	Shall document the choice and justify the relevance of evaluation metrics and their calculation methods Shall document each new evaluation metric
Shall be able to identify influence parameters and potential biases regarding operation conditions. Subgroups of functional specification and/or risk analysis shall be evaluated and performances measures shall be available per subgroup.	Take in account Preliminary Risk Analysis and functional specifications to divide the system in right subgroups.
Shall be able to identify if AI functionality is overlearnt or under-learnt.	Establish an evaluation protocol to detect that the AI is over/under learnt.
Shall evaluate Resilience and Robustness • Shall test in normal operating conditions • Shall test in degraded modes ○ Sensors failures (extreme values, out of range…) ○ Stress test or adversarial attacks	Behaviors observed out of operating domain of AI functionality shall be listed and documented and communicated to the clients. Qualification of test environments (virtual or real) shall be documented.
Shall evaluate the Reproducibility of the experiments and Repeatability of the performances. Ensure that variability of results is within given thresholds.	Elaborate a protocol that allows reproducibility and repeatability. The results of reproducibility and repeatability shall be traced. Variability in results difference must be within set thresholds defined during the design phase. Elaborate a protocol that allows to detect and trace cases of nonreproducibility and non-repeatability and must be communicated to the client.
Shall ensure that development/training and evaluation/validation teams are separated when preliminary risk analysis requires so.	Shall form the teams in accordance with preliminary risk analysis outcome.
Shall test the AI functionality with different user profiles.	

Evaluation Requirements	Requirements for Applicant
Shall test the AI functionality in real conditions. Performances deviations maximum level from controlled environment test shall not be crossed.	The choices of evaluation in real condition shall be documented.
Shall evaluate whether the results of evaluation comply with the conception phase requirements.	
Shall evaluate the compliance to regulation requirements of AI functionality	
Shall verify the risk analysis	The final risk analysis shall take in account development process and evaluation process feedback into account with regard to behaviors at limits or outside of the operational domain. This analysis must be documented and residual risk shall be communicated to the client.
Shall verify the documentation	If a company develops and uses its own IA functionality, the procession of decision making along with evaluation criteria must be documented

Table 8: Evaluation requirements and associated responsibilities of the applicant

From the design phase, a preliminary risk analysis (see requirement III.3.5), relevant and adapted to the use of the AI functionality, must identify, evaluate and document the risks associated with its use and their potential impacts.

The requirement precises that the risk analysis must foresee the use of erroneous data that may be due to sensor defects, formatting errors, bugs in the data management system or cyberattacks, and cover the components and sub-components as well as the interfaces between components of the AI functionality. The different failure modes of the AI functionality and their consequences must be established in order to allow the users to be aware of the residual risks they are exposed to.

The impacts can be quantified in terms of cost, safety, security, discrimination etc.

This preliminary risk analysis established during the design process (cf. III.3.5) shall be updated (see requirement III.5.11) taking into account the results of the development and evaluation processes of the AI functionality, in particular by taking into account the changes of behavior at the limits or outside the defined domain of use. This final risk analysis shall be documented and the residual risks shall be communicated to the customer.

5. Final conclusion

The need to ensure acceptable level of safety has been well identified and defined through AI-act [5] and other regulations on a European level. However, conventional methods of risk assessment are not applicable as is for AI enhanced application.

On the other hand certifying bodies have established protocols for certification and labellization of an AI based product even if, standards for means of compliance do not exist yet and are very complex to elaborate.

But, as this deliverable shows, many manufacturers and standard bodies are leading an effort on finding a way to assess and certify AI through many workgroup and conferences

Still as is, for manufacturers and standards bodies to find the right set of means of compliance, humans might have to stay in the loop and AI independent fall back systems should still be present to not compromise the safety. Moreover, As human, we cannot improve safety of an application by a product which itself needs to be proven safe.

6. Acronyms

Acronyms	Signification
AI	Artificial Intelligence
AIP	Actions Important for Protection
ASN	Autorité de Sûreté Nucléaire
ATM	Air Traffic Management
CAT	CATastrophic
CCA	Common Cause Analysis
CoDANN	Concepts of Design Assurance for Neural Networks
CSI	Common Safety Indicators
CSM	Common Safety Method
CSM RA	Common Safety Method for Risk Evaluation and Assessment
CSO	Common Safety Objectives
EASA	European Aeronautics Safety Agency
EIP	Elements Important for Protection
ERA	European Railway Agency
ERP	Emergency Response Plan
EU	European Union
FHA	Functional Hazard analysis
ForMULA	Formal Methods Use for Learning Assurance
FMEA	Failure Mode, Effects Analysis
FTA	Fault Tree Analysis
GAME	Globalement Au Moins Equivalent
GoA	Grades of Automation
HACCP.	Hazard Analysis and Critical Control Point
HAZOP	Hazard and operability analysis
HCTINS	High Committee for the transparence and Information of Nuclear Safety
HIC	Human In Control
HILEG	High Level Expert Group
HITL	Humans In The Loop
HOTL	Human On The Loop
ICAO	International Civil Aviation Organization
INB	Basic Nuclear Installation
IPL	Independent Protection Layer.
IRNS	Institute of Radioprotection and Nuclear Safety
KSF	Key Safety Function
LIC	Local Information Commissions
LOPA	Layer Of Protection Analysis
MLEAP	Machine LEarning application fro APplication
MOC	Maintenance in Operational Conditions
MOSAR	Méthode Organisée Systématique d'Analyse de Risques

Acronyms	**Signification**
NLF	New Legal Framework
PHA	Preliminary Hazard Analysis
PHA	Preliminary Hazard Analysis
PL	Performance Level
PRA	Preliminary Risk Analysis
PSA	Preliminary Safety Assessment
QRA	Quantitative Risk Assessment
R&D	Research & Development
RSD	Register of Dangerous Situations
SAE	Society of Automotive Engineers
SIL	Safety Integrity Levels
SIL	Safety Integrity Level
SMEs	Subject Matter Experts
SMS	Safety Management System
SOTIF	Safety Of The Intended Function
SSAs	System Safety Assessment
THR	Tolerable Hazard Rate
TSI	Technical Specifications for Interoperability
UE	Undesired Event
UML	Unified Modelling Language
UTM	UAS Traffic Management

7. List of Figures

8. List of Tables

9. References

[1] EU, Common Safety Method for Risk Evaluation and Assessment, 2009

[2] SAE, Guidelines for Conducting the Safety Assessment Process on Civil Aircraft, Systems, and Equipment, ARP4761, 2023

[3] SAE, Guidelines for Development of Civil Aircraft and Systems, ARP4754, 2023

[4] B. Debray & O. Salvi, INERIS, ARAMIS Project: an integrated risk assessment methodology that answers the needs of various stakeholders, available online : https://www.witpress.com/Secure/elibrary/papers/SAFE05/SAFE05027FU.pdf

[5] EC, «AI Act», 2021

[6] EC, NLF, available online : https://single-market-economy.ec.europa.eu/single-market/goods/new-legislative-framework_en, 2008

[7] ISO/IEC, Information technology - Artificial intelligence - Assessment of machine learning classification performance, ISO/IEC TS 4213, 2022

[8] ISO/IEC, Information technology - Artificial intelligence - Management system, ISO/IEC42001, 2023

[9] ISO, Quality management systems – Requirements - ISO9001, 2015

[10] ISO/IEC, Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML), ISO/IEC 23053, 2022

[11] EASA, EASA Artificial Intelligence Roadmap 1.0, 2020

[12] EASA, EASA Artificial Intelligence Roadmap 2.0, 2023

[13] EASA, Concept Paper: First usable guidance for Level 1 machine learning applications, 2021

[14] EASA, Concepts of Design Assurance for Neural Networks (CoDANN), 2020

[15] EASA, Concepts of Design Assurance for Neural Networks (CoDANN) II, 2021

[16] EASA/Collins Aerospace, Report-Formal Methods use for Learning Assurance (ForMuLA)-Public extract, 2023

[17] EUROCONTROL, The Fly AI Report - Demystifying and accelerating AI in Aviation/ATM, 2020

[18] SAE, Process Standard for Development and Certification/Approval of Aeronautical Safety-Related Products Implementing AI, ARP6983, WIP

[19] ASTRAIL, D3.2 AUTOMATIC TRAIN OPERATIONS IMPLEMENTATION OPERATION CHARACTERISTICS AND TECHNOLOGIES FOR THE RA, 2019

[20] LNE/DEC/CITI/CH, Laboratoire National de métrologie et d'Essais (LNE), Référentiel de certification de processus pour l'IA, Révision 2.0, 2021

[21] R. Regnier, Laboratoire National de métrologie et d'Essais (LNE), Cartographie LNE pour la certification et labellisation, 2021